ISB International Society for Biocuration

1st UK-Local Biocuration Conference

Abstract Book

May 5-6 2022

EMBL-EBI South Building - Ground Floor

Wellcome Genome Campus, Hinxton, Cambridge, UK

Contents

Welcome

Dear Colleague,

We are delighted to welcome you to Hinxton for the 1st UK-Local Biocuration Conference.

For many delegates this will be the first in-person conference you have attended since the onset of the COVID-19 pandemic and like us are excited to finally be able to meet with colleagues to share their work, promote collaboration and foster the sense of community the biocuration field has. We have encouraged participation from academia, government, and industry interested in the methods and tools employed in the curation of biological data and tried to give you every opportunity to share your knowledge with one another and discuss the future of biocuration.

We take this opportunity to thank the following sponsors, without whose support this event could not have taken place:
Healx – Keynote Speaker Sponsor
SciBite and Eagle Genomics – Talk and Poster Prizes Sponsor
GigaScience Press and AstraZeneca – Refreshments Sponsor

The ISB is a non-profit organization for biocurators, developers, and researchers with an interest in biocuration. The society promotes the field of biocuration and provides a forum for information exchange through meetings and workshops.

We hope that by attending this meeting you too will feel welcomed into our biocuration community.

Best wishes,
1st UK-Local Biocuration Conference Organising Committee

Organising Committee

Yasmin Alam-Faruque (Co-chair)	Healx, UK
Rachael Huntley (Co-chair)	SciBite, UK
Ruth Lovering	UCL, UK
Sandra Orchard	EMBL-EBI, UK
Arzu Ozturk Colak	FlyBase, University of Cambridge, UK
Pablo Porras Millan	AstraZeneca, UK
Mary-Ann Tuli	GigaSciencePress, UK

General Information

Conference Badges

Please wear your name badges at all times to promote networking and to assist staff in identifying you.

Internet Access

Wifi Access: eduroam or free guest Wifi (WGC Guest Wifi)

Social Media Policy

To encourage the open communication of science and biocuration, we would like to support the use of social media at this year's conference. Please use the conference hashtag #**UKbiocuration2022**. For poster sessions, please check with the presenter to obtain permission for sharing their work.

Poster Sessions

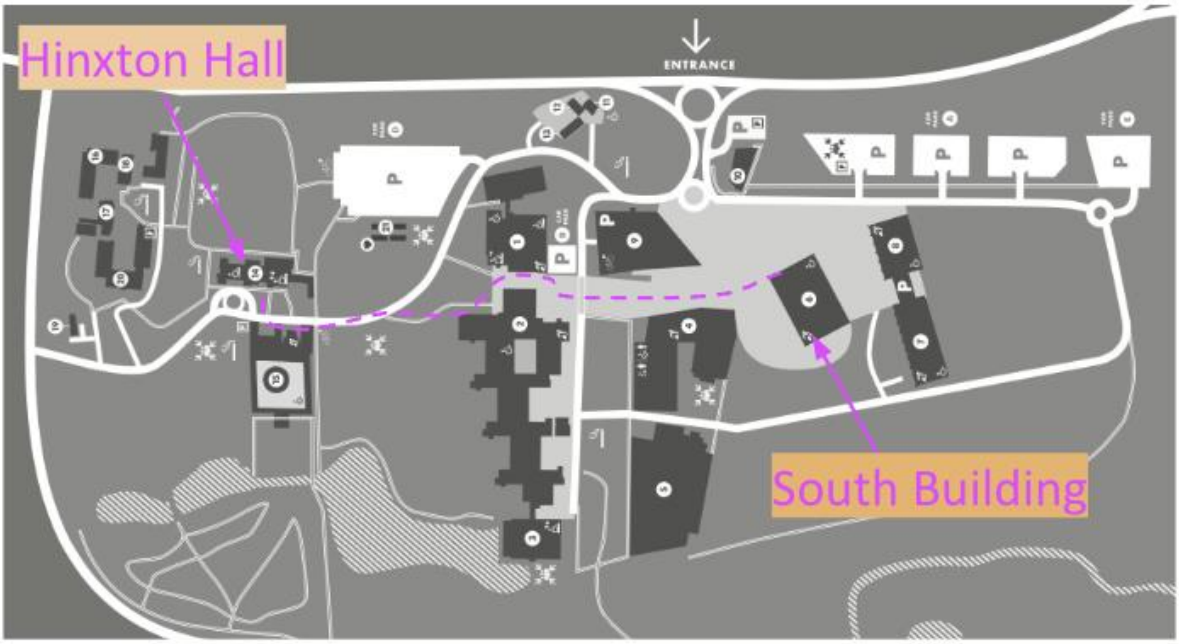
Poster sessions and lunch will be held at 12:40 on Thursday the 5th and at 12:30 Friday the 6th May. Posters numbered 1-12 will be presented on Thursday the 5th and posters numbered 14-25 will be presented on Friday the 6th. Poster abstracts are available on the pages 45-67 of the abstract book.

Networking session

The Hinxton Hall bar (please see the map below) will be open from 17:30 on Thursday the 5th, for anyone wanting to network or catch up with friends before the conference dinner.

Conference Dinner

There will be a buffet conference dinner at Hinxton Hall (please see the map below) on Thursday the 5th, starting at 18:30.



Photo/Video policy
Please note that photographs and footage will be taken throughout the event. These will be used by the ISB to promote future Biocuration meetings. Please use this form if you do not wish to appear in any photos or footage.

Taxis
Please find a list of local taxi numbers below:
Panther Taxis - www.panthertaxis.co.uk +44 (0) 1223 715715
CamCab - http://camcab.co.uk/ +44 (0) 1223 704704
A1 Cabco - http://www.a1cabco.co.uk/ +44 (0) 1223 313131
Cambridge City Taxis - https://www.cambridgecitytaxis.co.uk/ +44 (0) 1223 832832

Conference Sponsors

We are proud to be sponsored by the following organisations:

https://healx.io/

https://academic.oup.com/gigascience

https://www.astrazeneca.co.uk/

https://www.scibite.com/

https://www.eaglegenomics.com/

Conference Programme

Day 1 - Thursday 5th May

All events in Kendrew Lecture Theatre unless noted otherwise

09:00 **Registration**

09:30 **Opening & Welcome**

09:35 **Keynote Lecture**
Chair: Rachael Huntley
Connecting data to accelerate disease research
Mélanie Courtot

10:10 **Data Standards and Ontologies (FAIR) Session**
Chair: Rachael Huntley

10:10 **Dataset publishing to improve journal article transparency and reproducibility**
Christopher Hunter

10:25 **The role of validation in data curation - a Human Cell Atlas Data Coordination Platform case study**
Enrique Sapena Ventura

10:40 **FAIR Dataset Maturity**
Ibrahim Emam

10:55 **Challenges in standardization of Omics Data: Our Perspective**
Sanjanaa Jeevandass

11:10 Tea/coffee break

11:30 **Community Curation Session I**
Chair: Sandra Orchard
APICURON: attribution, quantification and real-time tracking of biocuration activity to promote engagement
Damiano Piovesan

11:45 **Software, Applications and Systems in Biocuration Session I**
Chair: Sandra Orchard

11:45 **SequenceServer 2.0: Improving BLAST visualization and analysis for unpublished or proprietary data**
Carlo Kroll

12:00 **Using Open Data and Reproducible Methodologies to Create an Global Inventory of Data Resources**
Heidi Imker

12:15 **ProtVar: Protein Coding Variant Annotation**
James Stephenson

12:30 **In Remembrance of Tony Sawford**

12:40 Lunch & Poster Session

12:40 **Poster Session**
Data Standards and Ontologies (FAIR) Posters
Software, Applications and Systems in Biocuration Posters

14:00 **Parallel workshops** *(Kendrew Lecture Theatre & Training rooms)*
1. Defining biocuration activities (Damien Piovesan - Apicuron)
2. FAIR and Industry (Rachael Huntley - SciBite)
3. Equity, Diversity and Inclusion (Mary-Ann Tuli - ISB)
16:00 Tea/coffee break
16:20 **Software, Applications and Systems in Biocuration Session II**
Chair: Sandra Orchard
16:20 **Towards Effortless Navigation of Scientific-Literature Screen Reading for Biocuration**
Sucheta Ghosh (Presenting via Zoom)
16:35 **Semantic digitization of experimental data: from information retrieval to biological discovery**
Pratibha Gour (Presenting via Zoom)
16:50 **Identifying clinically relevant biomarkers for paediatric cancers using text mining**
Jake Lever
17:05 **Wikidata Bib: a system for biocuration with analytics based on Wikidata**
Tiago Lubiana
17:20 **Updates from ISB**
Chair: Sandra Orchard
International Society for Biocuration (ISB): mission and goals in 2022
Federica Quaglia
17:30 Networking *(Hinxton Hall)*
18:30 Conference dinner *(Hinxton Hall)*

Day 2 - Friday 6th May

09:30 **Welcome**

09:35 **Keynote Lecture**

Chair: Yasmin Alam-Faruque

Preventing death from critical Covid using biological data: the Outbreak Data Analysis Platform (ODAP)

Kenneth Baillie

10:10 **Community Curation Session II**

Chair: Arzu Ozturk Colak

10:10 **BiCIKL – A community effort towards connected molecular, natural history collections, taxonomic data and literature**

Joana Pauperio

10:25 **JaponicusDB, creating a community sustainable Model Organism Database**

Valerie Wood

10:40 Tea/coffee break

11:05 **Data Curation Session I**

Chair: Ruth Lovering

11:05 **PRIDE Resources: Data deposition and dissemination**

Deepti Jaiswal Kundu

11:20 **More than the sum of its parts? Capturing, presenting and utilising gene product functional information at FlyBase**

Helen Attrill

11:35 **Refining the UniProtKB human proteome**

Michele Magrane

11:50 **Rfam, curating families with 3D structures**

Nancy Ontiveros

12:05 **Decoding Factors Inducing Perturbations in Biomolecular Networks: Data from the IMEx Consortium**

Kalpana Panneerselvam

12:30 Lunch & Poster Session

12:30 **Poster Session**

Community Curation Posters

Data Curation Posters

14:00 **Parallel workshops** *(Kendrew Lecture Theatre & Training rooms)*

4. Curate your career (Arzu Ozturk-Colak - ISB)

5. ELIXIR training resource collection (Alexandra Holinski - EBI Training Team)

16:00 Tea/coffee break

16:20 **Data Curation Session II**

Chair: Arzu Ozturk Colak

16:20 **Organ Anatomograms in Single Cell Expression Atlas**

Silvie Fexova

16:35 **Gene Ontology annotation of microRNAs**
Ruth Lovering

16:50 **The Global Biodata Coalition: Update and Progress**
Chuck Cook

17:05 **Equality, Diversity and Inclusion at ISB**
Chair: Arzu Ozturk Colak

A presentation of the work of the ISB's Equity, Diversity and Inclusion Committee
Mary-Ann Tuli

17:15 **Poster and presentation prizes**

17:30 **Conference close**

Keynote Speakers

Mélanie Courtot, PhD
Director of Genome Informatics
Ontario Institute for Cancer Research, OICR

Mélanie Courtot is Director of Genome Informatics and incoming Principal Investigator at the Ontario Institute for Cancer Research (OICR). Her team develops new software, databases and other necessary components to store, organize and compute over the large and complex datasets being generated by OICR's cancer research programs. Dr Courtot is passionate about translational informatics - building intelligent systems to gain new insights and impact human health. She co-leads the Data Use and Cohort representation groups for the Global Alliance for Genomics and Health (GA4GH), as well as cohort harmonization efforts for Common Infrastructure for National Cohorts in Europe, Canada, and Africa (CINECA), the International HundredK+ Cohorts Consortium (IHCC) and the Davos Alzheimer's Collaborative.

Prior to joining OICR in January 2022, Dr Courtot was the metadata standards coordinator for the archival and infrastructure team at the European Bioinformatics Institute (EMBL-EBI), where she designed tools to streamline multi-omics submissions and develops integrated metadata strategies across the institute's archival resources and other projects such as FAIRPlus, focusing on data quality, semantic enrichment, and standardization for pharmaceutical and cohort data respectively.

After receiving a BSc in Biochemistry and Master in Computer Science (2002) from the Université Louis Pasteur, in Strasbourg, France, Mélanie spent several years in different countries working as an international consultant/software developer. She rejoined academia in 2009 to start her PhD in Bioinformatics (graduated 2014) from the University of British Columbia, and did postdoctoral research at Simon Fraser University, before joining EMBL-EBI in June 2015 to lead the Gene Ontology (GO) editorial office and the Gene Ontology Annotation (GOA) projects.

Mélanie can be found on twitter, @mcourtot, where she often posts about science, equity and diversity, food and silly things she or her children do.

Kenneth Baillie, MD PhD
Professor of Experimental Medicine
Roslin Institute, University of Edinburgh

Kenneth Baillie graduated from the University of Edinburgh with a BSc(Hons) in Physiology in 1999 and MBChB in 2002. He completed basic training in medicine in Glasgow, and in anesthesia in Edinburgh. During this time he led a series of high altitude research projects in Bolivia, and founded a high-altitude research charity, Apex. He was appointed as a clinical lecturer on the ECAT (Edinburgh Clinical Academic Track) at the University of Edinburgh in 2008, and completed a Wellcome Trust-funded PhD in statistical genetics in 2012. He was awarded a Wellcome-Beit Prize Intermediate Clinical Fellowship in 2013. He led a global consensus on harmonisation of research studies in outbreaks for the International Severe Acute Respiratory Infection Consortium (ISARIC), and worked with WHO on H1N1 influenza, MERS, and Ebola. After completing clinical training in 2014 he worked as a visiting scientist at the Broad Institute of Harvard and MIT, before returning to the Roslin Institute, University of Edinburgh to establish a research program in translational applications of genomics in critical care medicine. He works as a consultant in the intensive care unit at the Royal Infirmary, Edinburgh. During the Covid outbreak in 2020-21, he led the UK-wide GenOMICC and ISARIC4C studies, and contributed to the design and delivery of the RECOVERY trial. He discovered new biological mechanisms underlying critical illness in Covid, and contributed to the discovery of effective drug treatments to reduce mortality.

He leads a research programme in translational genomics - using genetic signals from critically ill patients to identify both the targets for drug therapy, and the groups of patients likely to benefit most from any treatment, and testing those therapeutic ideas in highly-efficient model systems.

Talks

Data Standards and Ontologies (FAIR) Session

1. Dataset publishing to improve journal article transparency and reproducibility

Christopher Hunter, MA Tuli, CJ Armit, P Li, R Ménagé, K Cho, N Nogoy, SC Edmunds, L Goodman
GigaScience Press

With the ever-growing emphasis on FAIR principles since they were published in 2016, we are seeing more and more authors putting the effort into sharing their data. These efforts are often frustrated by the lack of specific guidance from journals and editors on exactly which data to share and how to make it FAIR. Some journals are trying to address this including both of our journals, GigaScience and GigaByte. In 2012 we created GigaDB in an effort to deliver the highest standards of data availability, transparency and reproducibility for all GigaScience Press publications. We take great pride in not only being the first publisher to offer in-house linked data hosting, but also with the fact that we employ specialist data curators to assist and advice our authors in the process that has become known as "FAIRification" of their work.

Here I will outline the steps we follow to curate a GigaDB dataset that is directly linked to a manuscript being published in GigaScience or GigaByte journals.

2. The role of validation in data curation - a Human Cell Atlas Data Coordination Platform case study

Enrique Sapena Ventura
EMBL-EBI

The Human Cell Atlas Data Coordination Platform (HCA-DCP) was established to ensure that the data from the HCA community is shared across the globe, and successfully utilised for the creation of a comprehensive map of the human body, one cell at a time.

In order to achieve this ambitious goal, one of the first steps for the data curation team was to determine what metadata is needed and how data needs to be structured. In the DCP, metadata is governed by the HCA Metadata Schema, a JSON schema that determines the requirements to ensure FAIRness of the stored projects.

The HCA metadata schema is an extensive collection of the entities that the HCA Data Coordination Platform supports. One of the reasons why a JSON schema was chosen over other options is the extensive JSON schema support, with libraries for validation in almost every coding language and an ability to support vast types of entries in a structured, yet human readable, way.

Alongside the HCA metadata standard, the HCA DCP also provides a flexible data model, which can support a variety of experimental designs. But with great power comes great responsibility, so these experimental designs need to be checked in order to ensure that certain constraints are observed. To solve these issues, we have implemented graph validation, which ensures that relational rules can be applied to the [meta]data.

All these checks and validation rules not only improve the quality of the data in the database, but also help accelerate the process of data curation and impact on the quality of service that users at both ends (Submitter and consumer) receive.

Ultimately, here we share our experience regarding data and metadata validation, past and current challenges and how implementing different levels of validation is important when thinking about building a highly curated database.

3. FAIR Dataset Maturity

Ibrahim Emam
Data Science Institute, Imperial College London; IMI FAIRplus

The FAIR principles were established in 2016 and have seen widespread adoption across the life sciences, being seen as a set of guidelines to promote good data management and stewardship. However, despite wide community adoption of the principles themselves, practical details on how to implement the principles to become FAIR are often too generic and lack the level of domain-specificity that facilitates actual execution of practical steps towards achieving real value-added FAIRified data. Within FAIRplus, we have used the FAIR principles to develop a practical and systematic framework for data FAIRification. Our approach consists of a set of general processes that can be adopted by any life-science-based data-generating projects, a FAIR maturity model for establishing the current state of FAIRness and a target state with practical domain-specific requirements to achieve the desired level of FAIR maturity. This talk will introduce the FAIR Dataset Maturity model, which aims to serve as a guidance for data generators in the life-science domain to systematically and un-ambiguously achieve value-added FAIR maturity for different data usage goals.

4. Challenges in standardization of Omics Data: Our Perspective

Sanjanaa Jeevandass, N Kurbatova, RB Mehta
Eagle Genomics, OBO, NCBI

Background: Biological data are omnipresent nowadays. They are available in different repositories, but we unable to retrieve the information efficiently as per our need, the field of omics is no exception. Omics data are spread across hundreds of open-source platforms, each one storing specific scientific information in their own prescribed format with limited integration with other sources. The struggle that scientists face day in and out is to get a holistic understanding of the available information on genomics, transcriptomics, proteomics, metabolomics and metagenomics (GenBank, NCBI, GO, etc.). Over the last few decades, as an industry, we have been proud digital revolutionists: attempting to move from paper to digital platforms aiming to solve the problem of accessibility and information sharing. The real question now however is: ”Are we able to leverage it to our advantage without spending much effort on processing, remapping, formatting the data-i.e. wrangling it?”. To do so, we must harmonize multiple datasets which is not only a cumbersome process but also time consuming and very specialized due to the knowledge required. So we see a common result, rather than researchers spending time on research, a huge portion of it is spent on collecting, curating, and pooling the data from various sources. This again emphasizes on the need for standardization, a key aspect of FAIRification. The main goal is interoperability between multiple data sources to extract additional biological knowledge that cannot be gained from a single dataset or omics modality alone. Merely pooling all the information into one database wouldn’t suffice since data context and meaning must be matched accurately. However, considering the complexity and depth of variety on omics data and the speed at which new data types are emerging, there is currently no good, standardized way for connecting various data sources and keeping them visible in one single platform.

Objective: For data standardization, the solution that we have tried and tested is an intermediate data model template that helps to transform data to a standardized structure. This template is modelled in alignment with various biological concepts and can comfortably ingest data from different data sources without losing biological context. Arriving at this template was an iterative process, and we faced a plethora of challenges ranging from physical versus conceptual entities to canonical sequence and protein isoforms. Highlighting these challenges to the audience and how we overcame them would help other scientists' standardization journey as for any individual/organization, a single fully integrated database containing all omics information is a dream come true and something that harnesses a lot of power through insights. Zifo have developed a common data model and has been able to formulate a successful path for data standardization. We see our model helping to move close towards FAIRification of omics data.

5. APICURON: attribution, quantification and real-time tracking of biocuration activity to promote engagement

Damiano Piovesan, Silvio C. E. Tosatto
University of Padova, Italy

Biocuration plays a key role in making research data available to the scientific community in a standardized way. Despite its importance, the contribution and effort of biocurators is extremely difficult to attribute and quantify. APICURON (https://apicuron.org) is a web server that provides biological databases and organizations with a real-time automatic tracking system of biocuration activities. APICURON calculates achievements and allows objective evaluation of the volume and quality of the contributions. Registered resources submit biocuration events to the APICURON web server. APICURON stores biocuration activities and calculates achievements (medals, badges) and leaderboards on the fly. Results are served through a public API and available through the APICURON website. APICURON is working with ORCID to automatically propagate badges and achievements to ORCID profiles. APICURON database schema is extremely simple and lightweight, however the definition of a hierarchical vocabulary of terms to describe all possible curation activities is work in progress. ELIXIR is already funding a project to support APICURON standardization and the integration of a core of early adopter’s curation databases (Pfam, Rfam, IntAct, SABIO-RK, PomBase, Reactome, SILVA and BioModels). APICURON aims at promoting engagement and certifying biocuration CVs. The whole community of biocurators is welcome to provide feedback and participate in all APICURON outreach activities.

Software, Applications and Systems in Biocuration Session I

6. SequenceServer 2.0: Improving BLAST visualization and analysis for unpublished or proprietary data

Carlo Kroll, DA Pava, A Priyam, M Alexandrakis, the SequenceServer community, Y Wurm
Queen Mary University of London, Wurmlab

BLAST analysis underpins many efforts for biocuration and research on the vast amounts of nucleotide and protein-sequences that are now available.

We developed SequenceServer to provide a point-and-click web interface for curators and researchers needing to analyse custom, unpublished, or proprietary data, or who want to avoid the queues and limitations of public web interfaces. SequenceServer is free and open-source and is used in thousands of labs, communities and companies, including for topics related to evolutionary analysis of newly sequenced genomes, cancer, neglected and emerging diseases, microbial genomes, vaccine development, and crop breeding.

The SequenceServer user interface is designed with a focus on improving the efficiency and abilities of curators and researchers to get their jobs done. The software benefits from using modern JavaScript libraries for creating interactive graphics.

Here, we will highlight recent and upcoming improvements to SequenceServer. We have improved the software's architecture to improve its robustness and customizability. It is now also easier to export and share analysis results. Furthermore, we have added several new visualization approaches that facilitate pairwise and many-to-many comparisons of genes and regions of interest, identifying orthologues, and identifying potential problems with gene predictions. These improvements should further help curators and researchers.

7. Using Open Data and Reproducible Methodologies to Create an Global Inventory of Data Resources

Heidi J Imker, Kenneth E Schackart, Ana-Maria Istrate, Chuck E Cook
Global Biodata Coalition, Chan Zuckerberg Initiative (CZI)

In the life sciences, biodata resources advance research by aggregating, organizing, preserving, and making accessible data to others, often free of charge and with no restrictions. With motivation and access to infrastructure, data resources can be created by anyone, anywhere, and at any time. While this means a relatively low bar to entry, maintaining these resources over time is a major challenge for individuals, funders, and the research enterprise as a whole. In order to better understand the challenge, we first must have a better understanding of the global infrastructure. The Global Biodata Coalition (GBC, https://globalbiodata.org/), in collaboration with the Chan Zuckerberg Initiative (CZI, https://chanzuckerberg.com/), is creating a comprehensive inventory of biodata resources to gauge the size of the global infrastructure. A key feature of the project is the development of reproducible methodology so that the inventory can be repeated at regular intervals. Our strategy relies on building a curated dataset by retrieving metadata from EuropePMC's API and building machine learning models that learn to identify mentions of biodata resources from the training corpus. We can then use the trained models on new text, to capture data resources not represented in current registries such as re3data.org and fairsharing.org. We employ state-of-the-art natural language processing techniques, such as pertained language models, which are openly available through the HuggingFace API. Additionally, by using articles and their associated clean and comprehensive metadata, we have the ability to gather more information about a data resource (such as a corresponding author email address, geolocation, associated article citations, etc.). This talk will outline our strategies for the project, including development of reproducible workflows, current results, and future directions.

8. ProtVar: Protein Coding Variant Annotation

James D Stephenson, R Ishtiaq, A Mishra, MJ Martin
Protein Function Development, EMBL-EBI

The consequences of variation in coding regions are predominantly driven by altering the wild-type behaviour of proteins. The two principal ways to understand if, and how, variants may affect protein function is to 1) compare with other residue co-located variants and 2) to understand the normal functional role of the amino acid in that protein position. However, collating the necessary information to investigate variants in this way is complicated by the variety of variant databases available, the fact that variants are described in terms of genomic location, the presence of alternative isoforms, and non standard numbering in protein structural models. ProtVar has been developed in order to contextualise protein coding variation in terms of protein function, structure and previously reported perturbations.

ProtVar uses the newly developed UniProt Variants Database which contains data from over 30 sources and then utilises newly pre-computed mappings between genomic positions and isoforms to retrieve position in UniProt canonical isoforms. The latest functional and structural data from UniProt and PDBeKB are then found for individual residues, functional regions and the protein in biological context. Through the ProtVar interactive user interface, user submitted lists of variants can then be assessed in a 3D protein structural and functional context which includes mutational experiments and previously reported variants at the same residue. The ProtVar API also allows programmatic access to the same functional data via genomic coordinates or amino acid positions with filtering of the results.

By integrating genomic variation with functional annotations ProtVar helps users to explore the potential impact of protein coding variation through a maintained, updated and tailorable platform.

Software, Applications and Systems in Biocuration Session II

9. Towards Effortless Navigation of Scientific-Literature Screen Reading for Biocuration

Sucheta Ghosh, W Mueller, U Wittig, M. Rey
Heidelberg Institute of Theoretical Studies (HITS gGmbH)

Background: the exponential growth of bio-literature has stirred the need for digitization of the biocuration processes. In this context, it is observed that screen reading of electronic documents paces up the whole process and also induces collaborative annotation in biocuration[1]. The usability experiences of screen-reading in the biocuration context are still an under-explored area. In other contexts, studies found that cumbersome navigational issues hamper intensive reading. It results in lower reading speeds (and thus, slower understanding) and fatigue when reading for an extended period, which leads to an increment in curation cost. On the other side, studies found that improvement of the organization of the electronic document can facilitate faster reading and understanding[2].

Objective: To our knowledge, this is the first and preliminary usability experience study to observe the effect of the navigation through the electronic document structure for the biocuration task.

Method: We conduct a curation task with the five curators for eighty papers in front of an eye tracker. We collect the level of difficulty on the Likert scale from the participants and their preferences for document structure. We explore different features for our statistical and correlation analysis: errors made by curators (checked by the other two expert curators), efforts made by the curators measured using an eye tracker. We used several eye tracker features, namely, time spent for reading, time spent for navigation, top-down reading time, bottom-up reading time, reading speed, fixation duration at different parts of the document, and pupillary responses.

Result: We find that the concurrent rhetorical and document structure facilitates fast reading. In this case, the navigational effort comes down significantly, especially for the experienced curators. It implies from this study that concurrent rhetorical structure[3] and document layout facilitate reading for digital biocuration. While in these experiment we did not see direct signs for explicit structural weaknesses of the paper, we are planning to investigate the impact of structural changes on the readiblity of papers. The key difficulty in this work is obtaining sufficiently large group of test persons able to read complex scientific publications.

[1] W3C Web Annotation Working Group. (2022, April 4). To enable a conversation over the world's knowledge: Hypothesis Mission. https://web.hypothes.is

[2] Hornbæk, K., & Frøkjær, E. (2003). Reading patterns and usability in visualizations of electronic documents. ACM Transactions on Computer-Human Interaction (TOCHI), 10(2), 119-149.

[3] Teufel, S., Siddharthan, A., & Batchelor, C. (2009, August). Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In Proceedings of the 2009 conference on empirical methods in natural language processing (pp. 1493-1502).

10. Semantic digitization of experimental data: from information retrieval to biological discovery

Pratibha Gour, Shaji V Joseph, Saurabh Raghuvanshi
University of Delhi

The advantages of digitizing and integrating experimental data in a semantic manner are many folds. Presently, the lack of standard storage/representation formats for the data published in research articles (which is by virtue highly validated, conclusive and thus invaluable) and immense heterogeneity among those datasets hinder their integration into systems biology models. In view of this problem, Manually Curated Database of Rice Proteins (MCDRP) attempted to curate the literature published experimental data itself. The data curation workflow in MCDRP utilizes in-house developed curation models, quintessentially based on universally accepted elements such as Plant Ontology, Environment Ontology, Trait Ontology, Gene Ontology and some other standard notations (e.g. RGAP gene locus id, GenBank protein id, NCBI taxon id, etc.). The varying aspects of a gene/protein and the smallest of the details of an experiment are broken down into small logical units and digitized in MCDRP using various existing ontologies and some custom-made alphanumeric notations. Such semantic curation potentially imparts a structure to the experimental data rendering it amenable to computer-based search and analysis. The biggest advantage it confers is the ease with which data can be searched and retrieved across literature. Every data descriptor that was used to curate the data can now be used in the different sections of the query builder to retrieve information instantaneously. Moreover, since the models enable digitization using unified elements and concepts, there is a natural and seamless connectivity conferred to the datasets. Consequently, experimental data published in numerous articles could be integrated and biological networks drawn such as those connecting genes/proteins based on their common spatial/temporal expression, responsiveness to common growth/environment condition, same trait regulatory behaviour, common sub-cellular localization, common biological process or common molecular function. All these correlations would be highly significant since the associations are based on validated experimental data. The creation of precise and comprehensive knowledge nests (via such literature curation tasks), giving an updated representation of existing biological information will provide researchers with novel insights into rice biology along with an easy access to highly granular and cross-linked information that can be traced all the way back to its source. This kind of an exercise in fact presents ‘meta- interpretation’ of literature i.e. analyzing multiple studies and interpreting over and above what a single study implies. Additionally, the use of universally acceptable ontologies and standard notations in these curation models makes the published experimental data and metadata considerably more FAIR (findable, accessible, interoperable, and reusable). This in turn enhances the chances of data-driven biological discovery.

11. Identifying clinically relevant biomarkers for paediatric cancers using text mining

Jake Lever, Jason Saliba, Obi L Griffith, Malachi Griffith and the ClinGen Pediatric Cancer Curation Advancement Subcommittee
University of Glasgow, Washington University School of Medicine

Knowledge of the underlying genetics of a patient’s tumour is revolutionising cancer treatment and empowering clinicians to make informed decisions about diagnosis, prognosis, risk and treatment options. Most cancer variant interpretation knowledge bases have targeted adult cancers. However paediatric cancers can have drastically different genetic landscapes and knowledge bases are desperately needed to provide good coverage of paediatric cancer specific information. As part of the ClinGen Pediatric Cancer Curation Advancement Subcommittee (PCCAS), we are enhancing the text mining system used for the CIViC knowledge base to improve coverage for paediatric information.

Our CIViCmine resource has provided text mining tools to assist in the curation of the CIViC cancer database. It extracts high-confidence sentences discussing mutations and their clinical relevance from abstracts and full-text papers. These sentences are then aggregated to identify frequently discussed biomarkers that should be manually reviewed for curation. We are extending this resource to highlight paediatric specific information.

To identify paediatric-specific papers, we have built tools to estimate the age demographics of different cancer types. This information along with paper MeSH information and journal information can be used to filter the CIViCmine dataset to highlight paediatric specific information. This text mined information will be used to identify paediatric knowledge already in CIViC that needs highlighting and prioritise papers for curation into the CIViC knowledge base.

12. Wikidata Bib: a system for biocuration with analytics based on Wikidata

Tiago Lubiana, Helder Nakaya
University of São Paulo (TL); Ronin Institute (TL); Hospital Israelita Albert Einstein (HN); Scientific Platform Pasteur-University of São Paulo(HN)

Biocuration often entails the systematic reading of a large number of scientific articles. Current tools for literature management (e.g. Mendeley and Zotero) are centered on PDFs, with little control over notes and virtually no possibilities to analyze one's reading history. Here we present Wikidata Bib, a reading framework based on scholarly information on Wikidata. The system, centered on Markdown, uses a GitHub repository for keeping notes under version control. A series of python scripts handle the reading flow and record statistics on articles read. The statistics feed a Wikidata-powered dashboard displaying meta information, like the most read authors and a map of their affiliations (available at https://lubianat.github.io/wikidata_bib/). We present the Wikidata Bib system alongside one use case: extracting cell types to catalog on Wikidata. We show that, besides analytics, the framework enabled a sizeable throughput, leading to over 1000 new cell types added to the knowledge base.

Updates from ISB

13. International Society for Biocuration (ISB): mission and goals in 2022

Federica Quaglia[1,2], Rama Balakrishnan[3], Parul Gupta[4], Robin Haw[5], Ruth Lovering[6], Sushma Naithani[4], Mary Ann Tuli[7], Randi Vita[8], and Nicole Vasilevsky[9]

[1]National Research Council (CNR-IBIOM), Bari, Italy; [2]Department of Biomedical Sciences, University of Padova, Padova, Italy; [3]Genentech, USA; [4]Department of Botany & Plant Pathology, Oregon State University, Oregon, USA; [5]Ontario Institute for Cancer Research, Toronto, ON M5G0A3, Canada; [6]Functional Gene Annotation, Institute of Cardiovascular Science, University College London, UK; [7]GigaScience Journal, Hong Kong, CN; [8]La Jolla Institute for Immunology, La Jolla, California, USA; [9]Translational and Integrative Sciences Lab, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

The International Society for Biocuration (ISB) is a professional society founded in 2009 to promote the field of biocuration and provide a community forum for information exchange and networking. The main goals of ISB include promoting best practices and career development, documentation and training on biocuration, providing awards recognising biocurator achievements and funding opportunities for ISB members.

The Society is overseen by an elected Executive Committee (EC) composed of nine members that in 2022 are: Nicole Vasilevsky (chair), Ruth Lovering (secretary), Robin Haw (treasurer), Rama Balakrishnan, Randi Vita, Mary Ann Tuli, Federica Quaglia, Sushma Naithani and Parul Gupta. There are also formal memberships in the society and anybody in the community is welcome to participate in most of the activities.

As the society has become more mature, the EC created specific subcommittees. The ISB has 7 subcommittees: Equality, Diversity and Inclusion (EDI), IT Infrastructure, Training, Outreach and Communication, Election Committee, Awards Committee, Travel/Exchange Fellowship Committee and Conference coordination committee. The work associated with these subcommittees is distributed amongst the EC members and provides ISB members with an opportunity to contribute to the aims of the ISB. More information about our subcommittees is available on our website.

Over the past 14 years, the ISB has hosted annual international conferences, entirely dedicated to the field of biocuration, with the location rotating between the three major regions of the world (the Americas, Europe, and Asia/Australia). These meetings bring together biocurators from various roles, including database curators, bioinformaticians, ontology developers and students. In 2021, the conference ISB2021, was held virtually in the form of four sessions and one workshop over the course of the year. Similarly, ISB2022 will feature three sessions during 2022 and we plan to resume in person conferences in 2023.

The ISB supports career development in biocuration, through dedicated pages on the ISB website that provide useful training materials, volunteering and career opportunities.

Benefits of becoming ISB members include, but are not limited to: i) supporting the work of the ISB, ii) financial benefits, e.g. discounted registration at ISB conferences, microgrants and fellowships, discounts on publication costs with Database journal

(Oxford), iii) opportunities to build networks with biocurators and the scientific community, e.g. Slack workspace, and iv) eligibility to serve the biocuration community.

ISB stakeholders span basic, clinical, and translational research communities and we proactively lead outreach activities through our website, Twitter channel (@biocurator), and quarterly newsletter.

Community Curation Session II

14. BiCIKL – A community effort towards connected molecular, natural history collections, taxonomic data and literature

Joana Pauperio, V. Balavenkataraman Kadhirvelu, V. Gupta, J. Burgin, S. Jayathilaka, A. Zirk, K. Abarenkov, K. Põldmaa, U. Kõljalg, J. Lanfear, L.Penev, Q. Groom, M. Dillen, A. Guntsch, D. Agosti, G. Cochrane

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom; Natural History Museum, University of Tartu, Vanemuise 46, Tartu 51014, Estonia; ELIXIR Europe, Cambridgeshire, United Kingdom; Pensoft Publishers, Sofia, Bulgaria; Meise Botanic Garden, Meise, Belgium; Botanic Garden and Botanical Museum Berlin, Freie Universität Berlin, Berlin, Germany; Plazi, Bern, Switzerland

The Biodiversity Community Integrated Knowledge Library (BiCIKL, https://bicikl-project.eu/) is a Horizon Europe project that aims to establish open science practices in the biodiversity domain, by providing FAIR access, developing tools and delivering services for linking data along the biodiversity research cycle (including specimens, sequences, species, analytics, publications and supporting re-use). BiCIKL joins together 14 partners and 15 research infrastructures that are working towards the provision of new methods and workflows for linking data from the data resources of molecular biology, natural history collections, taxonomy, and literature. Taking full advantage of these workflows requires a foundation of well-structured and accessible annotations, namely in the molecular sequence databases.

The molecular data maintained at the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/) hold a large amount of annotations relating to the sample source for an organism, as a natural history collection, a biobank or a culture collection. However, for a number of sequence records these annotations may be incomplete (sequences not linked to their sources), ambiguous (may lead to multiple endpoints) or even inaccurate. To establish a foundational layer of data for connecting molecular, natural history collections and taxonomic data there is the need to profile the existing source annotations of samples and sequences in the molecular databases and to build user tools and workflows for driving accurate and complete reporting of annotations or to provide updates when necessary.

We will showcase the initial work on the project in which we explore sample source annotations of molecular records, seek to improve the capture and validation of these annotations and begin to improve the connectivity between molecular sequence and natural history, taxonomic and biodiversity data.

Overall, we expect the work developed in the frame of the BiCIKL project to provide novel access to FAIR data that is fundamental in biodiversity research.

15. JaponicusDB, creating a community sustainable Model Organism Database

Valerie Wood, S Oliferenko, K Rutherford
PomBase, U of C; The Crick Institute; PomBase, U of C

The fission yeast *Schizosaccharomyces japonicus* has recently emerged as a powerful system for studying the evolution of essential cellular processes, drawing on similarities as well as key differences between *S. japonicus* and the related, well-established model *Schizosaccharomyces pombe*. We have deployed the open-source, modular code and tools originally developed for PomBase, the *S. pombe* model organism database (MOD), to create JaponicusDB (www.japonicusdb.org), a new MOD dedicated to *S. japonicus*. By providing a central resource with ready access to a growing body of experimental data, ontology-based curation, seamless browsing and querying, and the ability to integrate new data with existing knowledge, JaponicusDB supports fission yeast biologists to a far greater extent than any other source of *S. japonicus* data. JaponicusDB thus enables *S. japonicus* researchers to realize the full potential of studying a newly emerging model species and illustrates the widely applicable power and utility of harnessing reusable PomBase code to build a comprehensive, community-maintainable repository of species-relevant knowledge.

Data Curation Session I

16. PRIDE Resources: Data deposition and dissemination

Deepti Jaiswal Kundu, S Wang, A Prakash, S Hewapathirana, S Kamatchinathan, C Bandla, Y Perez, J A Vizcaino
European Molecular Biology Laboratory - European Bioinformatics Institute

Introduction-The PRoteomics IDEntifications(PRIDE) database at the EBI is currently the world-leading repository of mass spectrometry(MS)-based proteomics data. Thanks to the success of PRIDE and PX, the proteomics community is now widely embracing open data policies. Therefore, PRIDE has grown very significantly in recent years (~490 datasets per month were submitted on average during 2021). PRIDE has two main missions: (i) support data deposition and quality assessment of submitted proteomics experiments, to help reproducible research; (ii) promote and facilitate the re-use of public proteomics data, and disseminate proteomics evidences into added-value resources, including Ensembl, UniProt, MGnify and Expression Atlas. One major challenge is to ensure a fast and efficient data submission process, ensuring that the data representation is correct. Here, we describe in detail the PRIDE data handling and curation process.

Data handling and curation- A stand-alone submission tool is used during the data submission process. For each submitted dataset, an automatic validation is first performed to ensure that the data complies with the PRIDE metadata requirements and files in the dataset are correctly formatted and not corrupted during the upload process. Issues of different types can often be detected. At that point, the direct interaction between PRIDE curators and the users becomes critical. In a second step, the actual data submission takes place and accession numbers are provided to the users. Finally, datasets are publicly released when the corresponding paper is published. An ongoing challenge in PRIDE (and in the proteomics community as a whole) is to map the samples characteristics to the submitted files. The file format MAGE-TAB Proteomics has been recently developed as an extension of the original MAGE-TAB(used in transcriptomics) to enable this process. The submission of these files to PRIDE is currently optional.

Reanalysis and dissemination- Data re-use of public proteomics datasets has increased, with multiple applications. In this context, PRIDE datasets are routinely re-analyzed by the community. for sustainability reasons, our focus in-house has been put in disseminating and integrating PRIDE data into added-value EMBL-EBI resources such as UniProt, Ensembl, MGnify and Expression Atlas. The role of the curators in this process is to provide biological context, by manually curating the datasets that are reanalysed.

Conclusion- Data deposition and dissemination have changed the proteomics community since the creation of PX ten years ago. An increasing number of journals require nowadays the authors to deposit their data in a PX resource. The complexity of proteomics data makes a fully automated data deposition process very challenging, especially since data formats are complex and very heterogeneous. Curators then play a very active role in supporting the submitters in the preparation and quality control of each PRIDE data submission.

17. More than the sum of its parts? Capturing, presenting and utilising gene product functional information at FlyBase

Helen Attrill, Giulia Antonazzo and Nick Brown
FlyBase, University of Cambridge

Research in *Drosophila* has been central to the discovery and elucidation of many key aspects of the cell biology, development and physiology of multicellular organisms. Signalling pathways are a very active area of fly research, resulting in the identification of many new players and the detailed molecular dissection of pathway interactions. Using signalling pathways as a central example, we show how we have utilized the systematic curation of gene function using the Gene Ontology (GO) to model a dynamic, scalable and up-to-date picture of research. We also illustrate how we have used GO annotations in computational projects to seed and facilitate further studies.

Overall, this presentation aims to show how systematic curation of research findings is vital to the progress of biological research. Furthermore, it highlights the value of feeding the curated data into computational analyses to aid research on *Drosophila* signalling pathways, as well as other areas of biology.

18. Refining the UniProtKB human proteome

Michele Magrane, Paul Denny, Yvonne Lussi, Sandra Orchard, UniProt Consortium
EMBL-EBI, SIB Swiss Institute of Bioinformatics, Protein Information Resource

UniProtKB provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. It facilitates scientific discovery by organising biological knowledge and enabling researchers to rapidly comprehend complex areas of biology. Access is provided to protein sets for species with sequenced genomes from across the tree of life. This includes provision of the complete human proteome which contains the set of protein sequences derived by translation of all protein-coding genes of the human reference genome. A first-pass curated human proteome was made available in UniProtKB in 2008 but its content is not static and continues to evolve as technologies advance and new findings emerge. Sequences are updated to reflect improvements in the underlying gene models and to align with efforts such as MANE (Matched Annotation from NCBI and EMBL-EBI). In addition, proteins are added or removed as new information about the coding status of the underlying genes arises, and the proteome continues to expand through the inclusion of novel isoforms produced through processes such as alternative splicing. Work also continues on improving the functional information associated with each protein to ensure that users have access to a high-quality comprehensive human protein set. All data are freely available from www.uniprot.org.

19. Rfam, curating families with 3D structures

Nancy Ontiveros-Palacios, E Cooke, A Bateman, B A. Sweeney and A I. Petrov
EMBL-EBI

Rfam (https://rfam.org) is a comprehensive database of RNA families, each represented by a manually curated alignment, a consensus secondary structure and a covariance model. Rfam is nearly 20 years old and since its inception, it has become the largest and most widely used database of non-coding RNA (ncRNA) families, currently including over 4,000 families. Historically, the RNA secondary structure for each family was obtained through bioinformatic predictions before a 3D structure was solved. Now, the Rfam team is using available 3D information from Protein Data Bank (PDB) to review ncRNA families and provide more accurate secondary structures. As of March 2022 the 3D structures of 124 Rfam families have been reported and we are currently updating the families by adding missing base pairs, as well as other missing structural elements, including pseudoknot elements and long distance base pairs. Additionally, we are including annotations, such as RNA 3D motifs (for example k-turn), RNA structural elements (stems, hairpin loops, junctions), and ligands (particularly in riboswitches). In the 30 first reviewed families, the secondary structure of 27 families was improved by updating missing pseudoknots in 19 of the families and adding or correcting base pairs in 26 of the families. For example, the central part of the flavin mononucleotide (FMN) riboswitch is now organized by several additional base pairs and two pseudoknots. The first 30 updated families have been released and include riboswitches, coronavirus RNAs, spliceosomal RNAs, ribozymes, microRNAs, and other RNAs. Rfam is continuously improving the quality of RNA families, and this targeted review of families using 3D information fills the gap between RNA-predicted structures and the experimentally determined RNA 3D structures.

20. Decoding Factors Inducing Perturbations in Biomolecular Networks: Data from the IMEx Consortium

Kalpana Panneerselvam[1], Pablo Porras[2], Noemí del-Toro[3], Margaret Duesbury[1], Livia Perfetto[4], Anjali Shrivastava[1], Sandra Orchard[1], Henning Hermjakob[1], IMEx Consortium Curators

[1]*European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI);Data Curation Director,* [2]*AstraZeneca;* [3]*Data Content Engineer, Healx;* [4]*SIGNOR database coordinator, Human Technopole*

Molecular interaction (MI) networks provide maps to explore cellular processes from a systems perspective. Understanding the network environment of a given biomolecule provides critical insight into the processes in which it is involved and the mechanisms by which it is regulated.

The IMEx consortium (www.imexconsortium.org) is an international database collaboration that exists to record molecular interaction data from scientific literature and direct depositions and make it freely accessible to the public, using an Open Access, Open-Source model.

IMEx curators follow a deep curation model representing experimental evidence presented in the scientific literature in great detail; so far more than 1.18 million binary interactions have been curated. The model allows curators to accurately document the factors that can introduce perturbation within the network. These contributing factors include mutations, variable experimental conditions, chemical and biological inhibitors, agonists and antagonists and required post-translational modifications.

About 18K binary interactions have been shown to be disrupted by a total of >74K single amino acid mutations or the deletion of required binding regions. The required binding region may not make direct contact with its partner but is identified as a region of a molecule being absolutely required for an interaction. Additionally we have annotated 1500 new binary interactions introduced by mutated proteins when compared with the wild-type, which does not interact. Around 75% of the mutation annotation are mapped to human proteins, providing high-quality experimental evidence of sequence change effects which directly relate to existing variation data. Additionally we have curated required PTMs that control over 1000 binary interactions; and information on known agonists and antagonists which affect ~10,000 and 2,500 binary interactions respectively.

The IMEx consortium has been generating contextual molecular networks that can serve as a scaffold for analysing the perturbations induced by these factors for our best understanding on requirements of protein interaction. This openly available resource is an invaluable tool with immediate applications in the study of variation impact on the interactome, sub-networks generated by mutated partners, small molecules affecting the networks, interaction interfaces as drug targets, among other key questions.

All data is available under a CC-BY licence from https://www.ebi.ac.uk/intact.

Data Curation Session II

21. Organ Anatomograms in Single Cell Expression Atlas

Silvie Fexova, Jana Eliasova, Lingyun Zhao, Nancy George, Alfonso Munos-Pomer Fuentes, Jonathan Manning, Pablo Moreno, Sarah Teichmann and Irene Papatheodorou
Gene Expression Team, EMBL-EBI; Cellular Genetics, Sanger Institute

Single Cell Expression Atlas is an open science resource for scRNA-Seq data, that enables gene queries across studies, cells, cell types and species. The resource has grown steadily since its launch in 2018 and now contains over 300 single-cell datasets and more 8.5 million cells from 20 different species. It disseminates data from a variety of major projects, including the Human Cell Atlas, Fly Cell Atlas, discovAIR and others.

The data in Single Cell Expression Atlas are presented in either a t-SNE or UMAP plots which showcase the variability of gene expression at the single cell level. However, it can be difficult for a user to fully relate data from t-SNE or UMAP plots and clusters to the real-life complexity of the biological tissues they represent and to see the cells and organs behind the dots. For this reason, we have developed a new interactive data visualisation tool – the organ anatomograms. The anatomogram is an anatomy diagram of a human organ or region within. It consists of a chain of interlinked, interactive images that display an organ and its substructures in increasing level of detail, all the way to the cellular level. Its individual component parts are annotated with ontology IDs and the anatomogram pipeline matches these with the inferred cell type annotations in each dataset and leverages the ontology structure to also highlight corresponding parent structures in any of the higher-level images within the given organ anatomogram stack. This puts individual cell types identified through analysis of single-cell sequencing experiments in broader structural context within each tissue/organ. Anatomograms also allow users to quickly discover top cell type markers for each cell type in an experiment and development is under way to connect them to other plots and tables within the experiment page as well as across experiments in Single Cell Expression Atlas. Currently, we have released anatomograms for lung, pancreas, placenta and liver with kidney and gut in production and more anatomograms (including anatomograms for other biological species) and more functionality linked to them on the way. The anatomogram components are flexible and can be easily embedded into different resources to display different types of human datasets.

22. Gene Ontology annotation of microRNAs

Ruth C Lovering, G Antonazzo, A Pesala, DL Buitrago, P. Asanitthong, M Long, M Makris, H. Attrill, C. Logie, P Gaudet

Functional Gene Annotation, UCL Institute of Cardiovascular Science, University College London, London; FlyBase, Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge; Molecular Biology Department, Faculty of Science, Radboud University, Nijmegen, The Netherlands; Swiss-Prot group, SIB Swiss Institute of Bioinformatics, Geneve, Switzerland

MicroRNAs are small, ~22-nucleotide molecules that silence gene expression by binding to complementary target mRNA strands. MicroRNA (miR) regulation of developmental and cellular processes is a relatively new field of study, however the data generated from such research needs to be optimally organised to allow inclusion of this data in pathway and network analysis tools. Although there are several miRNA bioinformatic databases and tools that include information on expression, sequence and target data, many of these rely on predicted targets or text mining.

The association of proteins with terms from the Gene Ontology (GO) has proven highly effective for large-scale analysis of functional data, therefore, we have applied this approach to describe the biological role and mRNA targets of miRs. There are several challenges associated with manual curation of miRs and it is important to ensure their consistent annotation.

In this poster we will present the revised GO Consortium guidelines for curation of miRs, and decision trees for deciding the terms to apply and how to confirm the species of the miR to curate. In addition, we will present how our annotations are contributing to the understanding of the role of miRs in co-ordinating the regulation of specific processes.

The annotations we have created are freely available in the QuickGO browser, RNAcentral, miRBase, FlyBase, Ensembl and the PSICQUIC web server, enabling our data to be included in many functional analysis tools and used to interpret high-throughput datasets.

Funded by National Institute for Health Research University College London Hospitals Biomedical Research Centre, the British Medical Research Council grant #MR/N030117/1 and the National Human Genome Research Institute (NHGRI) grant U41 HG002273.

23. The Global Biodata Coalition: Update and Progress

Chuck E Cook, R Drysdale
Global Biodata Coalition

Research in the life sciences is data-driven and dependent on the data integration and analysis enabled by open-access biodata resources. These resources form a major global infrastructure that is crucial to current and future biomedical and life sciences research, but the infrastructure has developed organically without systematic coordination among funders who spend hundreds of millions of dollars every year supporting the infrastructure, and there is no long-term strategy to ensure sustainability of this infrastructure. These resources continue to grow rapidly, as they are crucial to biomedical and life sciences research and are required to store the data generated by the open access policies adopted by most funders, yet there is no comprehensive understanding of the infrastructure at a global level.

The Global Biodata Coalition (GBC: https://globalbiodata.org/) is supported by research funders and aims to better coordinate and share approaches for the efficient management and growth of biodata resources worldwide. The GBC's missions are to help funders ensure sustainable financial support for the global biodata infrastructure and in particular to identify a set of Global Core Biodata Resources that are crucial for sustaining the broader biodata infrastructure.

To further these missions the GBC currently has three active projects:
1) The Funders Board has convened two working groups to take a global perspective and explore the implications of their policies on (i) biodata resource sustainability and (ii) open data strategies for biodata resources.

2) The global infrastructure of biodata resources has developed organically over many decades: consequently the size and scope of the ecosystem of biodata resources is not well-described. As an important step in defining this infrastructure the GBC is undertaking an inventory of global biodata resources that will give a first ever worldwide overview of biodata resources, providing a basis upon which the GBC can work with funders to coordinate support for biodata resources and for the entire infrastructure.

3) The GBC is also working to identify a set of Global Core Biodata Resources (GCBRs), defined as those biodata resources that are fundamental to the entire life science data infrastructure. The details of this process were published in January 2022 (https://zenodo.org/record/5845116) and Expressions of Interest were accepted from 21st March - 22nd April 2022
(https://globalbiodata.org/scientific-activities/gcbr-selection/).

In this talk we will provide updates on these three active projects.

Equality, Diversity and Inclusion at ISB

24. A presentation of the work of the ISB's Equity, Diversity and Inclusion Committee

Mary-Ann Tuli
GigaScience Journal, BGI-Hong Kong

The ISB's Equity, Diversity and Inclusion (EDI) committee was formed as an outcome of a workshop held at the 2019 ISB Biocuration conference in Cambridge, UK. Its goals are to promote Equity, Diversity and Inclusion across the Society and the broader community.

The committee, which meets once a month, is currently comprised of 8 ISB members, three of whom also sit on the Executive Committee.

We have many ways of disseminating our work and are mindful of being accessible to the very people we are attempting to reach.

Consequently, we have hosted workshops and webinars (both in-person and virtually), written publications, maintained an EDI page on the ISB website and have contacted biocurators in underrepresented groups. Topics we have discussed include the use of alcohol in professional events, the gender balance at ISB meetings, career progression, professional code of conduct & how to reach out to colleagues in underrepresented countries.

This presentation will allow me to go into more detail about our work and to communicate its important message.

Workshops

1. Defining biocuration activities

Damiano Piovesan, Henning Hermjakob, Silvio Tosatto
University of Padova, EBI

Biocuration plays a key role in making research data available to the scientific community in a standardized way. Despite its importance, the contribution and effort of biocurators is extremely difficult to attribute and quantify.

Several efforts have been made during the last decade to credit the work of biocurators in manually curated resources. One obvious solution is to use the Open Researcher & Contributor ID (ORCID, https://orcid.org/). ORCID allows integration of ‘Works’ such as publications, data sets, conference presentations, etc. in the user profile page. However, ORCID is not designed to directly manage biocuration activities and it does not provide a system to aggregate and weigh this type of information. APICURON (https://apicuron.org/) is a new service aiming at filling this gap by providing aggregated statistics, rankings and featuring real-time tracking of biocuration activities. APICURON calculates achievements and allows objective evaluation of the volume and quality of the contributions. However, it is not yet integrated with ORCID and lacks standardization.

Indeed the definition of a curation activity is itself problematic. The World Wide Web Consortium (W3C) Provenance Working Group already defined generic entity activities such as “generation”, “invalidation” and “revision”. However, curation activities in different resources can be extremely different and can be tracked at different levels of granularity. Moreover, it is extremely difficult to quantify the curation effort for all possible activities.

In order to standardize curation activities and provide curation databases with a structure to capture this information coherently it is necessary to have a broader discussion bringing biocurators, database developers and attribution service providers (ORCID and APICURON) together around the same table.

2. FAIR Data and Ontologies in Industry

Rachael Huntley, Jane Lomax
SciBite

Abstract: The use of ontologies in the pharmaceutical industry is becoming increasingly commonplace. Ontologies are essential for companies to ensure they can easily organise, search and access legacy data, as well as enable consistent metadata at point-of-entry. Increasingly, it is a priority for the industry to adhere to FAIR data capture to build a FAIR data architecture that can be implemented enterprise wide. This effort relies heavily on the provision of high-quality public ontologies, which can be built upon and tailored to specific use cases within each company. In this workshop we will bring together representatives from both the pharmaceutical industry and companies that provide curation, ontology and thought-leadership services to them. Each speaker will give a 15 minute presentation to illustrate the importance of FAIR data in their companies and how ontologies are used to achieve their goals. We will follow this with a 40 minute panel session where we will encourage the audience to provide questions and discussion points.

Bios of the confirmed speakers:

George Georghiou (Novartis)

George received his PhD in Molecular and Cellular Pharmacology from Stony Brook University focusing on the structural biology of small molecule inhibition of protein kinases. He then went on to a post-doc at the MRC Protein Phosphorylation and Ubiquitylation Unit at the University of Dundee focusing on the use of CRISPR/Cas9 in studying protein kinase family signaling. Afterwards, George joined the Gene Ontology Annotation and UniProtKB group at EMBL-EBI as a data curator, responsible for annotating targets for the CAFA project and later providing cross-referencing between UniProtKB disease terms to the Experimental Factor Ontology for OpenTargets. After 3 years at the EBI, he joined Novartis as an information scientist in the Novartis Knowledge Center, where he currently creates custom data solutions as well as FAIRifying the data the Knowledge Center licenses from a variety of vendors.

Peter McQuilton (GSK)

Peter works in data quality and governance, focusing on ontology management, information architecture and FAIR metadata. He holds a PhD in Genetics and NeuroDevelopment from the University of Cambridge and has published over 30 scientific papers. Working as a Biocurator before it was cool, Peter started out as literature curator at FlyBase, before moving to the University of Oxford to lead the domain-agnostic FAIRsharing resource. After 20 years in academia, working on data and metadata curation, data management, and creating the odd ontology or two, Peter is now responsible for the Ontology Management Platform, working on Information Architecture and FAIR at GSK. He enjoys thinking about all the ways one can store, reconcile, maintain and serve semantically rich FAIR knowledge on an enterprise scale.

Shirin Saverimuttu (SciBite)

Shirin joined SciBite in December 2021 as a Scientific Curator. She is involved with developing ontologies for customers as well as updating SciBite's pre-existing vocabularies. Prior to starting at SciBite, Shirin spent just over a year working at EMBL-

EBI as a curator for both the Polygenic Score and GWAS Catalogs. Her background includes a BSc in Human Biosciences and a MSc in the Genetics of Human Disease.

Stacy Mather (AstraZeneca)

Stacy is the Head of FAIR Data Services in the Data Office at AstraZeneca. She has built the capability and led a dedicated team through rapid growth and new service delivery since 2020, while improving cycle times from weeks to days. As the Head of FAIR Data Services she is responsible for end-to-end service delivery for FAIR-ification of R&D data including data asset management, data curation, data provisioning and reference & master data management.

Prior to joining AZ, Stacy has held several senior leadership positions at a major global CRO in the technology services division, mainly in operational service delivery. She is originally from the US, and had several international posts including scaling up operational delivery in the APAC region, with assignments in Tokyo and Shanghai. Stacy now resides in the UK with her family and very active lab x lurcher cross.

3. Equity, Diversity and Inclusion

Mary Ann Tuli
GigaScience Press, BGI-Hong Kong

Attendees of ISB conferences have now come to anticipate that there will be an EDI workshop during the meeting and so we are excited to have the opportunity to host what is always an insightful and lively session.

A growing number of institutes now have dedicated EDI staff, which support and advise employees, management and human resource teams. This recognises that EDI issues are valuable, complex and acknowledged.

With invited speakers and a panel comprising EDI experts and scientists the workshop will both educate and invite debate and discussion.

This workshop is aimed at all delegates: students, people at the beginning of their careers and more senior and experienced scientists.

4. Curate your career: what your future career could look like and how to get it

Arzu Ozturk Colak, Ruth Lovering, Rachael Huntley
FlyBase - University of Cambridge, UCL, SciBite

Abstract: There are a variety of different biocuration or biocuration-related careers available. During this workshop there will be 15 minutes presentations from biocurators and managers from a variety of working environments. The speakers will be encouraged to discuss some of the advantages and disadvantages of different biocuration roles and what biocurators can do to improve their future careers. Audience will be encouraged to raise questions to expand on these topics during the panel discussion (40 minutes).

Bios of the chair and speakers:
Pablo Porras Millan (AstraZeneca) (Chair)
Pablo Porras started his career in the University of Córdoba, Spain, focused on redox homeostasis, then moved to a Neuroproteomics group of the Max Delbrueck Center, Berlin, where he developed his interest in interactomics. During this period, he faced the problem of how to represent and analyze complex interaction and OMICs data, an experience that led him to join EMBL-EBI to work as a scientific curator, bioinformatician and project coordinator in IntAct. At EMBL-EBI he got involved in projects dealing with FAIR, metadata representation, literature information extraction, analysis of interaction data sets, bioinformatics resource development, plus ontology and standards management and creation, with a strong mandate on servicing scientific research. In May 2021, Pablo joined AstraZeneca as Data Curation Director within the R&D Data Office, an exciting new role focused on providing reusable data representation practices and know-how, following FAIR principles, to the clinical research and development field.ion practices and know-how, following FAIR principles, to the clinical research and development field.

Sameer Velankar (EMBL-EBI)
Sameer runs the Protein Data Bank in Europe (PDBe) team which manages three resources. As a member of the wwPDB consortium, he is involved in managing the single global archive of macromolecular structure data the Protein Data Bank (PDB). His team also manages the macromolecular structure knowledge base (PDBe-KB) and the AlphaFold Protein Structure Database. Sameer completed his PhD from the Indian Institute of Science Bangalore, India. He then moved to Oxford before joining the EMBL-EBI in 2000. Sameer was keen to make structural data accessible to the wider scientific community, and to facilitate research and has worked over the last 20 years on building infrastructure and tools for managing and analysing macromolecular structures.

Ruth Lovering (University College London)
Ruth has been involved in annotation of the human genome for 22 years and now leads an annotation team based at UCL, which has submitted around 9% of the manual Gene Ontology annotations associated with human genes. In addition, this team has supported UCL staff with data analysis or data interpretation. Ruth also has a variety of teaching roles including running a bioinformatics core MSc module and an annual two-day bioinformatics workshops at UCL, which have enabled over 400 scientists to learn how to use a variety of online genomic resources in their data analyses and interpretation.

Jane Lomax (SciBite)

Jane leads the development of SciBite's vocabularies and ontology services. With a PhD in Genetics from Cambridge University and 15 years' experience working with biomedical ontologies, including at the EBI and Sanger Institute, where she focussed on bioinformatics and developing biomedical ontologies. She has published over 35 scientific papers, mainly in the area of ontology development, and contributed to public ontology projects including Pistoia Ontology Mapping Project and was on the Executive Committee for the ISB.

Eleanor Stanley (Eagle Genomics)

Eleanor Stanley started her biocuration career in 1993, before the role was formerly named, working first for FlyBase in University of Cambridge and after 5 fantastic years moving to UniProt to extend her invertebrate knowledge from genetic, phenotypic and ecological to protein and biochemical aspects. At UniProt, bioinformatics skills were encouraged that resulted in an MSc(Res) qualification and a position at the Wellcome Trust Sanger Institute building a genome annotation platform for the 50 Helminth project. The move to industry in 2014, taking a position with Eagle Genomics, was with some trepidation but was absolutely the right thing to do - curation skills supporting her current role as Data Governance manager, with a healthy splash of information security too.

5. ELIXIR training resource collection
(Presentation of an online training collection that meets the training needs of some biocurators and those who want to become one)

Alexandra Holinski, Anna Swan, Sarah Morgan
Training Team, EMBL-EBI

In 2019 ELIXIR launched a global study[1] with the aim of gaining an up-to-date profile of biocuration. This study aimed to identify communities of biocurators, the type of biocuration work being done, training needs and gaps, and to draw a picture of biocuration career development. In this study, programming / scripting / coding, development and use of ontologies, and database management were identified as the most common training needs. In addition, skills like programming and extracting data from literature were mentioned as vital for a successful career in biocuration. The study also showed that there were few training offerings available to meet the specific training needs that biocurators have. To address this we will run a webinar series throughout March and April 2022 targeted at biocurators and addressing topics like programming skills, an introduction to text mining, the use of ontologies and biocuration careers in industry. The recordings of these webinars will become part of an online training collection for biocurators comprising additional tutorials on data management, user experience research, and examples of biocuration efforts.

In this workshop we will:
- Present the online training collection on biocuration
- Ask attendees to revise the collection and provide feedback on topics they wish to have extended or that are missing
- Discuss ways in which biocurators can be supported in satisfying their training needs
- Encourage attendees to share experiences and advice on how to learn new skills

The workshop will also be an opportunity for attendees to get involved in shaping the further development of the training collection and in actively contributing material to it.

[1] Holinski A, Burke ML, Morgan SL et al. Biocuration - mapping resources and needs [version 2; peer review: 2 approved]. F1000Research 2020, 9(ELIXIR):1094 (https://doi.org/10.12688/f1000research.25413.2)

Posters

Data Standards and Ontologies (FAIR)

1. The *Drosophila* Anatomy Ontology

Clare Pilgrim, Rob Court, Damien Goutte-Gattat, Alex McLachlan, David Osumi-Sutherland
University of Cambridge; University of Edinburgh; EBI

The *Drosophila* Anatomy Ontology (DAO) is a queryable store of knowledge about *Drosophila* anatomy and cell types. It consists of over 18,000 terms and over 50,000 formal assertions curated from almost 1,500 publications. It supports the multiple classification schemes used by biologists, for example classifying neurons by lineage, neurotransmitter, sensory modality and innervation pattern.

Building and maintaining a high-quality ontology of this complexity would not be possible without automation. Consistency and efficiency are achieved by using standard design patterns and scripting. Many new neuron types have recently been identified from electron microscopy data and thousands of these have now been added to the DAO in a standardised way.

Terms are created and updated in a collaboration between FlyBase and Virtual Fly Brain (VFB), both of which use the DAO to annotate data. VFB takes advantage of the logical assertions recorded in the DAO to drive queries, for example allowing users to query for neurons by innervation pattern.

We will present examples of how we use automation to aid ontology development and of how the knowledge encoded in the DAO is used to drive biologically meaningful queries on VFB.

2. Should data fields in biological resources be labelled more FAIRly?

Bryony Braschi, Liora Vilmovsky , Tamsin Jones, Ruth Seal, Susan Tweedie and Elspeth Bruford

HGNC, EMBL-EBI, University of Cambridge

The HUGO Gene Nomenclature Committee (HGNC, genenames.org) is responsible for approving unique symbols and names for human loci, including protein coding genes, ncRNA genes and pseudogenes, to allow unambiguous scientific communication. We also name genes in selected vertebrates via our sister project, the Vertebrate Gene Nomenclature Committee (VGNC, vertebrate.genenames.org).

Approved nomenclature from our websites, together with curated alias symbols and names, are widely displayed by other resources including Ensembl, NCBI Gene, The Alliance of Genome Resources and UniProt, which aids FAIR data sharing. However, our data (and other data) are often labelled differently when they appear on another resource's website; what we call an "Approved symbol" may be labelled as a "Gene name", which may help to explain why many of our users confuse the terms "name" and "symbol". For other data types many different terms are used to refer to exactly the same data – is this causing confusion for users?

At a time when FAIR principles are encouraged and the use of ontologies is widespread is it worth devoting more effort to standardising labels for data fields? Many users will be interested in gathering data from many sites across many species - but having to learn a new vocabulary for each site and work out what fields are equivalent isn't ideal. We will highlight the differences in data labels used across several key biological resources, and explore the pros and cons for a more standardised approach. We feel some standardisation could help users to navigate between resources and are keen to discuss how this could be achieved, or learn of efforts that are working to make this happen.

3. ShareYourCloning: First steps towards a FAIR standard for cloning strategies

Manuel Lera-Ramirez
UCL

ShareYourCloning is a web application to generate molecular cloning strategies in json format. The aim of this application is to provide a web interface and an underlying data model to document the generation of new DNA molecules from existing ones, and to export this information to share it with others.

The data model contains two kinds of entities:

- Sequences: which represent sequences and their features (using .gb format).
- Sources: which represent the origin of sequences, which may be external sources (e.g. the ID of a plasmid that was received from Addgene), or experimental sources, representing cloning steps combining existing entities to generate new entities.

Currently, the application uses a custom data model in json to be able to keep a lean web application development, but in the future we would like to adopt SBOL (Synthetic Biology Open Language).

For more info, visit: https://github.com/manulera/ShareYourCloning

5. The role of validation in data curation - a Human Cell Atlas Data Coordination Platform case study

Enrique Sapena Ventura
EMBL-EBI

The Human Cell Atlas Data Coordination Platform (HCA-DCP) was established to ensure that the data from the HCA community is shared across the globe, and successfully utilised for the creation of a comprehensive map of the human body, one cell at a time.

In order to achieve this ambitious goal, one of the first steps for the data curation team was to determine what metadata is needed and how data needs to be structured. In the DCP, metadata is governed by the HCA Metadata Schema, a JSON schema that determines the requirements to ensure FAIRness of the stored projects.

The HCA metadata schema is an extensive collection of the entities that the HCA Data Coordination Platform supports. One of the reasons why a JSON schema was chosen over other options is the extensive JSON schema support, with libraries for validation in almost every coding language and an ability to support vast types of entries in a structured, yet human readable, way.

Alongside the HCA metadata standard, the HCA DCP also provides a flexible data model, which can support a variety of experimental designs. But with great power comes great responsibility, so these experimental designs need to be checked in order to ensure that certain constraints are observed. To solve these issues, we have implemented graph validation, which ensures that relational rules can be applied to the [meta]data.

All these checks and validation rules not only improve the quality of the data in the database, but also help accelerate the process of data curation and impact on the quality of service that users at both ends (Submitter and consumer) receive.

Ultimately, here we share our experience regarding data and metadata validation, past and current challenges and how implementing different levels of validation is important when thinking about building a highly curated database.

6. The power of combining and expanding drug target synonyms for target safety assessment

Lucy Sheppard
In Silico Solutions, Instem

For the evaluation of potential drug safety issues, PubMed is a key information resource as it provides the greatest overview of general scientific activity relating to the biology of a drug target (i.e. protein or gene). For the purpose of building a corpus of scientific literature pertaining to a drug target, target synonyms can be gathered from various sources, including nomenclature resources (e.g. HUGO Gene Nomenclature Committee), genomics knowledgebases (e.g. Ensembl, UniProt) and pharmacology databases (e.g. ChEMBL), or extracted from the literature itself (e.g. PubTator). To further maximise recall, a list of target synonyms may also be semi-automatically expanded, e.g. adding or replacing punctuation, and using spelling variations. By combining and expanding target synonyms from various sources for each member of the novel protein kinase C family (PRKCD, PRKCE, PRKCH and PRKCQ), we were able to increase the corpus size of PubMed abstracts 1.5- to 25-fold (depending on the original synonym source).

We chose to further examine the corpora for PRKCQ (protein kinase C theta), which is of active pharmaceutical R&D interest for the treatment of autoimmune diseases and cancer. Using automated tagging of vocabulary around drug safety hazards, the larger "meta" PRKCQ corpus contained potential associations with 1.5- to 3-fold more hazards than the largest (PubTator) or smallest (UniProt) corpora built using unexpanded target synonyms from one source. We manually curated the potential associations of PRKCQ with hazards involving the hepatobiliary and nervous systems, which are two key areas of safety concern. Within the larger "meta" PRKCQ corpus, we validated the association of modulating PRKCQ function with 16 novel nervous system hazards (e.g. amnesia, paralysis and stroke) and 4 novel hepatobiliary hazards (e.g. hepatic steatosis and hepatotoxicity), which did not appear in the next-largest PRKCQ corpus derived from unexpanded PubTator synonyms. This demonstrates the power of combining and expanding official and literature-derived target synonyms for the purpose of assessing safety issues associated with modulating the function of drug targets. Instem's aim to continually improve the scope and quality target safety related data-sets provides actionable intelligence to the pharmaceutical industry.

7. Accelerated variant curation from scientific literature using text mining

R Mallick, V Arnaboldi, P Davis, S Diamantakis, A Becerra, Magdalena Zarowiecki , KL Howe

Genome Assembly and Annotation, EMBL-EBI; WormBase, Caltech

Biological databases collect and standardize data through biocuration. Even though major model organism databases have adopted some automation of curation methods, a large portion of biocuration is still performed manually. To speed up the extraction of genomic variant data, we have developed a hybrid approach that combines regular expressions, Named Entity Recognition based on BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) and bag-of-words to extract variant genomic locations from *C. elegans* papers for WormBase, along with associated gene, strain, variant name and species. Our model has a precision of 82.59% for the gene-mutation matches tested on 100 papers, and even recovers some data not discovered during manual curation. Code at: https://github.com/WormBase/genomic-info-from-papers

8. The Genomics Standards Consortium MIxS v6 release

Chris I Hunter, all GSC members
The Genomics Standards Consortium

The Genomic Standards Consortium (GSC) is an open-membership working body formed in September 2005. The aim of the GSC is to assist in making genomic data discoverable. The GSC enables genomic data integration, discovery and comparison through international community-driven standards.

Over the years we have assisted in the design and implementation of 11 minimum information checklists and over 20 supplementary environmental package checklists, together we refer to all of these as "MIxS" - Minimum Information about any(x) Sequence.

During the last 2 years we have made a massive effort to update this set of checklists and packages both in terms of their individual definitions and content as well as the technology used to maintain and disseminate those checklists and packages. These changes are bundled together in the most recent release, version 6, of the MIxS checklists. Here I will present some of the major changes in this release including; new environmental packages and an overview of the LinkML technology that is now being used in its implementation.

9. UniProt automatic annotation systems: UniRule and ARBA

Pedro Raposo, H Zellner
EMBL-EBI

UniProt is a well established database that offers comprehensive information about proteins to the scientific community. With the ever-growing number of genomes being sequenced and the increase of translated protein sequences, there is a need, more than ever, to automatically classify and provide functional information for proteins for which experimental information might never become available.

UniRule and ARBA are two of the automatic annotation systems UniProt uses to annotate more than one hundred million unreviewed UniProtKB/TrEMBL entries based on the presence of InterPro sequence signatures. Whilst both of these tools base their annotation prediction on reviewed UniProtKB/Swiss-Prot data, UniRule involves the manual creation of rules by curators, and ARBA uses association rule learning to automatically gather and process information. This work demonstrates how these systems provide detailed information in a reliable and accurate way, and how these can be used by users to annotate independent protein datasets.

10. SequenceServer 2.0: Improving BLAST visualization and analysis for unpublished or proprietary data

Diego A Pava, C Kroll, A Priyam, M Alexandrakis, the SequenceServer community, Y Wurm
Queen Mary University of London, Wurm Lab

BLAST analysis underpins many efforts for biocuration and research on the vast amounts of nucleotide and protein-sequences that are now available.

We developed SequenceServer to provide a point-and-click web interface for curators and researchers needing to analyse custom, unpublished, or proprietary data, or who want to avoid the queues and limitations of public web interfaces. SequenceServer is free and open-source and is used in thousands of labs, communities and companies, including for topics related to evolutionary analysis of newly sequenced genomes, cancer, neglected and emerging diseases, microbial genomes, vaccine development, and crop breeding.

The SequenceServer user interface is designed with a focus on improving the efficiency and abilities of curators and researchers to get their jobs done. The software benefits from using modern JavaScript libraries for creating interactive graphics.

Here, we will highlight recent and upcoming improvements to SequenceServer. We have improved the software's architecture to improve its robustness and customizability. It is now also easier to export and share analysis results. Furthermore, we have added several new visualization approaches that facilitate pairwise and many-to-many comparisons of genes and regions of interest, identifying orthologues, and identifying potential problems with gene predictions. These improvements should further help curators and researchers.

11. COSMIC’s Variant Pipeline, Rising to the Challenge

Amaia Sangrador-Vegas, CG Cole, A. Holmes, C Kok, S Wang, R Stefancsik, L Ponting, B Harsha, C Boutselakis, N Bindal-Dhir, T Maurel, SA Ward
COSMIC, Wellcome Sanger Institute

COSMIC, the Catalogue of Somatic Mutations in Cancer, is the largest and most comprehensive resource to explore the impact of somatic mutations in cancer. Cancer develops through the acquisition of a series of mutations that cause normal cells to transform into cancer cells, and the mutational landscape of tumours can be complex, with most mutations not directly involved in cancer.

In the last two decades, advances in next-generation sequencing (NGS) have made sequencing more accessible to researchers, which has revolutionised our understanding and treatment of cancer. There are two types of sequencing approaches in cancer samples. Targeted sequencing focuses on specific candidate genes or gene regions. Genome-wide applications aim to discover the broad landscape of genes/variants involved in tumour development, and can analyse the whole genome (WGS) or focus on the exons (WES).

Depending on the type and breadth of a study, publications can describe a limited number of variants or include hundreds of thousands. To accommodate these different scenarios, COSMIC’s pipeline facilitates both manual and programmatic curation of variants. Manually curated variants are added through an in-house interface using the cDNA and/or protein nomenclature for canonical transcripts. The systematic pipeline, ATRAC (Annotation TRAnsfer to COSMIC), was designed to add large amounts of data using genomic coordinates derived from VCF files from genome-wide screens and large targeted screens. All variants are stored in the Oracle COSI database. This data is subsequently combined and standardised for genome versions GRCh38 and GRCh37 using COSMIC’s DIAS (Data Integration and Annotation System). For all variants where the genomic location is known, DIAS uses Ensembl’s VEP (Variant Effect Predictor) to re-annotate variants for all transcripts to a specific Ensembl version following HGVS nomenclature. Redundant variants are merged and simple nucleotide variants are given stable COSMIC genomic identifiers (COSV). Furthermore, given the complexity of the data - which includes many different types of mutations and fusions - additional processing takes place at this stage, and the resulting information is stored in the Curation database, which is used to build the warehouses for each release.

COSMIC’s pipeline requires expert curation backed by powerful bioinformatic processing. Currently, processing all this data is time consuming and allows only 2 or 3 releases per year. We are working to develop a more efficient pipeline that will optimise COSMIC’s ability to deal with the continuous growth in genomic data and to better reflect the consequence of mutations (e.g., those causing exon skipping). As per release v.95, COSMIC describes 5 million coding mutations across cancer, close to 16 million non-coding mutations and 19,422 gene fusions.

12. A survey of phyletic profiling tools to bolster bioinformatic usability, interoperability, and reproducibility

Colbie J. Reed, Geoffrey Hutinet, Rémi Denise, Valérie de Crécy
The Department of Microbiology & Cell Science, University of Florida

A phylogenetic distribution or phyletic profile is often an essential component of bioinformatic analysis at the family level. Frequently, the use of phyletic profiling is intended to help resolve potential functional shifts within and between protein families by mapping the presence-absence of various sequence traits (e.g., fused domains, catalytic motifs), including gene duplications, to respective taxa, illustrating patterns of implicated functional conservation, divergence/convergence. This approach can also be applied at the scale of metabolic pathways to better resolve gene-losses, incidents of non-orthologous gene displacement (NODs). These observations can be essential grounds for further in silico hypothesis generation in bioinformatics, particularly for NOD candidate identification. Phylogenetic profiling tools rely on ortholog identification, and, therein, the results can be greatly affected if paralogs are not correctly identified. The diversity in size and complexity of protein families makes “cookie-cutter” tools prone to errors and, further, may require the development of custom tools and/or independent workflows. This notably contributes to the challenge of establishing a single platform that meets the needs of each and every bioinformatic investigation, no matter the family’s taxonomic distribution or number of paralogs/distinct protein fusions.

An ideal phylogenetic distribution analysis platform should allow for: 1) custom genome selection; 2) custom selection of multiple protein/gene families, orthologous groups; 3) customizable ortholog identification/filtration methods; 4) user-friendly graphic-data (vector) interfacing, transformation; 5) automated aggregation, visualization of data source-linked genome and sequence annotations; 6) interoperable, customizable data and figure export. Here, we compared available phylogenomic resources—primarily public-facing webtools—to evaluate how they fulfilled this set of ideal needs. It was found that to date no single web-tool combines all of these attributes. Some allowed genome selection but had poor family or orthologous groups. Others demonstrated good family selection tools but had little to no flexibility of genome selection. In all cases, the dependence on precomputed protein family tools or annotations was an issue that greatly restricted the utility of the platforms examined. Potential solutions are discussed.

14. Annotation Projects: A Student Perspective

Ruth C Lovering, A Deng, S de Miranda Pinheiro, K Thurlow
Functional Gene Annotation, UCL Institute of Cardiovascular Science, University College London, London

Bioinformatics data is exploited by pharmaceutical, biomedical, academic and industrial researchers, however, limited focus is on teaching this area to students and senior scientists. Consequently, many scientists develop their own expertise without appreciating the source of the data they are reliant upon. Some universities and institutes provide courses on a selection of bioinformatics resources and tools, and may also provide biocuration projects, during which students submit data to annotation resources. To assess whether students find biocuration projects useful and / or enjoyable, a survey was sent to University College London (UCL) students who have undertaken Gene Ontology annotation projects. Analysis of the responses confirmed that these projects were enjoyed by the students and that they appreciated the opportunity to learn about bioinformatics resources and to improve their literature analysis skills. After undertaking these projects, potentially these students will critically assess their own manuscripts and ensure that these are written with the biocurators of the future in mind.

This work was funded through grants from the Alzheimer's Research UK (ARUK-NAS2017A-1) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. We are extremely grateful to all of the students who contributed GO annotation and completed the survey.

Data Curation

15. COSMIC Cancer Gene Census and Hallmarks of Cancer: Identification and functional annotation of genes causally associated with cancer

Mike Starkey, J Wilding, CG Cole, SA Ward, Z Sondka
COSMIC, Wellcome Trust Sanger Institute

The identification and characterisation of genes that are involved in tumourigenesis is a prerequisite for the development of precision cancer therapeutics. In this regard, a challenge is that more than 1% of the genes in the human genome are implicated in one or more cancer types through somatic and/or germline alterations.

The Cancer Gene Census is a longstanding and ongoing manually curated resource established to consolidate research into the molecular pathogenesis of cancer by providing a compendium of genes affected by mutations that contribute to cancer development. Regular review to establish the eligibility of new genes for inclusion in the Census proceeds through consistent rigorous evaluation of an objective set of criteria. Central to this is the curation of multiple independent published research studies consistently demonstrating the frequent somatic alteration of a gene in one or more cancer types and/or the occurrence of rare pathogenic germline mutations associated with a familial cancer/cancer syndrome. The existence of a mutation profile consistent with a gene having one, or more, recognisable roles in one, or more cancer types, is an additional evaluated metric.

Demonstrating a causal association between a gene and one or more cancer types also requires experimental evidence of how a gene may contribute to tumour development and progression. The Hallmarks of Cancer are the phenotypic traits that delineate tumours from normal tissues and which characterise all cancer types. COSMIC has adapted the tumour-level hallmark behavioural characteristics to both provide a framework for evaluating the potential functional contribution to cancer development of candidate Census genes, and subsequently a means of functionally describing how Census genes are involved in tumour development and progression. Evidence is collated by the continuing manual curation of reproducible results from multiple independent published research studies in which in vitro assays, and/or experiments featuring animal models, have been deployed to evaluate the effects of individual genes on one or more cancer hallmarks. This evidence is the basis for the functional annotation of the proteins encoded by wild-type Census genes as promoting and/or suppressing each cancer hallmark in one, or more, cell type contexts.

Genes listed in the Cancer Gene Census are categorised into two tiers based on the strength of the evidence associating them with cancer. For Tier 2 genes, evidence of frequent mutations of a type and distribution that enable designation of a role in cancer, or of a functional contribution to cancer development, is less well established than for genes included in Census Tier 1.

The Cancer Gene Census released with COSMIC v95 catalogues 729 genes, including 6 recent additions. Hallmarks of Cancer functional annotations are provided for 318 Tier 1 genes, including new annotations for 15 genes.

16. Manual Functional Annotation of Transcription Factors from *C. elegans* in UniProtKB

Paul Denny, Rossana Zaru, Leonardo Jose da Costa Gonzales, Elena Speretta, Michele Magrane, Sandra Orchard, UniProt Consortium
UniProt, EMBL-EBI; Swiss Institute of Bioinformatics; Protein Information Resource

The UniProt database is a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. This includes proteins from all branches of the tree of life, but with an emphasis on humans and a number of key model organisms, such as the mouse, fruitfly and the nematode worm, *Caenorhabditis elegans* (*C.elegans*). We work with both the nematode worm research community and with WormBase, the database of the biology and genome of *C. elegans* and related nematode species, to ensure that UniProtKB presents detailed and current functional annotation of nematode proteins.

Transcription factors play a central role in regulating gene expression and so act as keystones for almost all biological processes. Our main focus in this project is the manual annotation of experimentally characterised transcription factors (TFs) from *C. elegans*. Curators add published experimental data to automatically-annotated TREMBL entries, prioritising breadth of coverage of TFs, primarily DNA-binding transcription factors (dbTFs), but also general (gTF) and co-regulatory (coTF) factors. As a part of this process, TFs were associated with specific Gene Ontology terms and where possible, regulated target genes were included. This work built upon the substantial efforts made to improve the transcription branches of the ontology undertaken by the Gene Ontology Consortium. Given suitable experimental data, the breadth of representation of the different classes of DNA-binding domains was augmented. Almost all known worm dbTF families now have at least one member manually curated. Another priority has been to choose to annotate TFs with mammalian orthologs, in order to increase the utility and re-use of UniProt data.
We will report on progress in expanding the manual annotation of transcript
ion factors encoded by the genome of the roundworm, *C. elegans*.

17. The pipeline integration and manual curation of proteomics data in UniProt.

Emily H Bowler-Barnett, J Fan, S Orchard
Protein Function Content, EMBL-EBI; Swiss Institute of Bioinformatics, Centre Medical Universitaire; Protein Information Resource, Georgetown University Medical Center

Proteomics analyses can provide a wealth of protein sequence, peptide modification, interaction, and protein function data. The addition of such large-scale studies into UniProt must be handled with care and attention. Accurate interpretation, analysis, and curation of such studies is of high importance and serves to enrich all aspects of protein understanding and annotation. Conclusions derived from mass spectrometry studies can be diverse and built upon in various ways, it is therefore important to annotate these outcomes in a way befitting their end points, both in methodology as well as in understanding.

Protein sequence data is imported into the UniProt database using a defined pipeline from collaborative databases, it then undergoes filtering, in-silico analysis and quality control before peptide sequence data is matched to UniProt protein sequences and entries. Post-translation modification analysis using mass spectrometry is becoming more widespread and understood by the proteomics community, integration of this data is challenging, hampered by complex analysis pipelines. Additionally, proteomics analyses can form part of multiplex studies encompassing peptide identification, differential protein abundance analysis, protein-protein interaction partner detection and even protein function discovery.

The annotation of protein sequence and proteomics-derived data enriches the UniProt database, allowing it to act as a repository for an ever-expanding set of reference proteomes and protein function annotation across the phylogenetic kingdoms. The import and manual annotation of proteomics data directly enriches the UniProt database at the protein entry level and is increasingly augmenting protein visualization and presentation of proteomics data.

All data are freely accessible from www.uniprot.org

18. Mapping from virulence to pathogen-host interactions; an annotation review

Antonia Lock, S Orchard
Protein Function Content, EMBL-EBI

Proteins in UniProt are assigned keywords from automated rules and manual curation. Some keywords are mapped to Gene Ontology (GO) terms (the KW2GO pipeline), so that proteins annotated to a specific keyword automatically are assigned associated GO terms.

The GO is continuously revised to reflect current understanding of concepts in biology, and terms may be obsoleted if they are deemed out of scope for GO. In March 2021 the term GO:0009405 pathogenesis was obsoleted as it does not describe a single normal Biological Process, rather it is the effect of an interaction between two organisms, under specific conditions.

Through KW2GO, the keyword KW-0843 Virulence had been mapped to the obsoleted GO:0009405 pathogenesis. As KW-0843 Virulence is associated with 4,114 manually reviewed (Swiss-Prot) entries and 147,664 unreviewed entries (Trembl), the loss of the mapping resulted in a loss of over 150,000 annotations.

A GO consortium review of manual annotations to GO:0009405 pathogenesis curated from the literature, resulted in the removal of almost 40% of annotations. The remaining 60% of annotations were re-annotated to the branch of GO describing interspecies interactions. The review of manual literature annotations to GO:0009405 pathogenesis prompted a review of entries annotated to KW-0843 Virulence, prior to remapping of the keyword to GO.

19. The modified GWAS Catalog: increasing data volume and improving data representation

Santhi Ramachandran, Ala Abid, Annalisa Buniello, Maria Cerezo, Peggy Hall, James Hayhurst, Arwa Ibrahim, Sajo John, Elizabeth Lewis, Abayomi Mosaku, Elliot Sollis, Fiona Cunningham, Lucia Hindorff, Laura Harris, Paul Flicek, Helen Parkinson
European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, United Kingdom; Open Targets, Cambridge, United Kingdom; Division of Genomic Medicine, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, United States

The NHGRI-EBI GWAS Catalog (www.ebi.ac.uk/gwas) is a repository of all published genome-wide association studies (GWAS) with >39,000 studies, including 28,459 studies with full genome-wide summary statistics and 347,165 top associations as of March 2022. In addition, the GWAS Catalog also includes around 1,151 pre-published summary statistics datasets in a standard format.

Publications that qualify for inclusion are identified weekly by a machine-learning-assisted literature search. Curators assess the publications identified by literature search to consistently extract traits, top significant associations, and sample meta-data. One of the challenges is to curate large and complex data and make it available in the Catalog within 1-2 months of the publication. In recent years, in fact, the Catalog's data volume has increased rapidly (~987% increase in no. of studies in Q1-2022 compared to Q1-2021), and this has required curation to scale up to maintain the rate of throughput. In addition, software improvements have significantly helped make manual data extraction less time-consuming. While all GWAS Catalog traits present are mapped to Experimental Factor Ontology (EFO) terms to enable standardisation of traits across all studies, representation of ancestry and cohorts is still challenging, as few established guidelines are available. The Catalog has therefore developed and published a framework to represent ancestry in a consistent way which has reduced missing data.

Summary statistics and metadata can be directly submitted by users via the web interface prior to publication. Curators then QC and annotate the data prior to release. This means faster turnaround time for publications and more data completeness. The summary statistics data sharing has now become more common. Currently, around 70% of GWAS studies in the Catalog have full summary statistics available; this shows that GWAS Catalog has now become an important part of the data-sharing community. Since there is no accepted community standard for GWAS data generation and sharing, we have worked with the community to develop a standard format based on the most commonly included fields in publicly available files to maintain consistency in the Catalog and facilitate submission by users.

20. Parasite proteomes: a challenge for protein annotation in UniProt Knowledgebase and gene ontology (GO)

Rossana Zaru, Michele Magrane, Sandra Orchard, and UniProt Consortium
EMBL-EBI, Cambridge, UK, Swiss Institute of Bioinformatics, Centre Medicale Universitaire, Geneva, Switzerland, Protein Information Resource, Georgetown University Medical Center, Washington, USA, Protein Information Resource, University of Delaware, Newark, USA

Parasites are either single cell organisms, such as the causative agent of malaria *P. falciparum*, or multicellular organisms including some fungi and worms, which have a substantial impact on human and livestock health and agriculture. Their complex life cycle, unique cellular structures and their genetic adaptability to various environments and host defences pose many challenges for the annotation of their proteomes. For instance, how to annotate the interaction between parasite and host/vector proteins, how to highlight strain polymorphism? How to report variation between host/vector and parasite homolog proteins? And eventually how to make this information easily searchable and retrievable?

The UniProt Knowledgebase (UniProtKB) collects and centralises functional information on proteins across a wide range of species. For each reviewed protein entries, we provide information about its function, interactors, cellular localization and expression as well as an extensive annotation of sequence features. We also provide information about sequence polymorphism and biotechnological use.

Using various examples from the malaria parasite *P.falciparum*, I will illustrate how, in UniProtKB, we have addressed some of the challenges in the annotation of parasite proteins. Ultimately, improving the curation of parasites and their hosts and vectors proteomes will help the scientific community to better understand parasite biology, their interaction with hosts and vectors, explore their phylogeny, identifying potential targets for drug development and designing strategies to manipulate vectors.

21. Decoding Factors Inducing Perturbations in Biomolecular Networks: Data from the IMEx Consortium

Kalpana Panneerselvam[1], Pablo Porras[2], Noemí del-Toro[3], Margaret Duesbury[1], Livia Perfetto[4], Anjali Shrivastava[1], Sandra Orchard[1], Henning Hermjakob[1], IMEx Consortium Curators

[1]*European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI);Data Curation Director,* [2]*AstraZeneca;* [3]*Data Content Engineer, Healx;* [4]*SIGNOR database coordinator, Human Technopole*

Molecular interaction (MI) networks provide maps to explore cellular processes from a systems perspective. Understanding the network environment of a given biomolecule provides critical insight into the processes in which it is involved and the mechanisms by which it is regulated.

The IMEx consortium (www.imexconsortium.org) is an international database collaboration that exists to record molecular interaction data from scientific literature and direct depositions and make it freely accessible to the public, using an Open Access, Open-Source model.

IMEx curators follow a deep curation model representing experimental evidence presented in the scientific literature in great detail; so far more than 1.18 million binary interactions have been curated. The model allows curators to accurately document the factors that can introduce perturbation within the network. These contributing factors include mutations, variable experimental conditions, chemical and biological inhibitors, agonists and antagonists and required post-translational modifications.

About 18K binary interactions have been shown to be disrupted by a total of >74K single amino acid mutations or the deletion of required binding regions. The required binding region may not make direct contact with its partner but is identified as a region of a molecule being absolutely required for an interaction. Additionally we have annotated 1500 new binary interactions introduced by mutated proteins when compared with the wild-type, which does not interact. Around 75% of the mutation annotation are mapped to human proteins, providing high-quality experimental evidence of sequence change effects which directly relate to existing variation data. Additionally we have curated required PTMs that control over 1000 binary interactions; and information on known agonists and antagonists which affect ~10,000 and 2,500 binary interactions respectively.

The IMEx consortium has been generating contextual molecular networks that can serve as a scaffold for analysing the perturbations induced by these factors for our best understanding on requirements of protein interaction. This openly available resource is an invaluable tool with immediate applications in the study of variation impact on the interactome, sub-networks generated by mutated partners, small molecules affecting the networks, interaction interfaces as drug targets, among other key questions.

All data is available under a CC-BY licence from https://www.ebi.ac.uk/intact.

22. From genotypes to phenotypes: Disease curation in UniProtKB

Yvonne C. Lussi, Elena Speretta, Kate Warner, Michele Magrane, Sandra Orchard and UniProt Consortium
Protein Function Content, EMBL-EBI; Swiss Institute of Bioinformatics SIB; Protein Information Resource PIR

The UniProt Knowledgebase (UniProtKB) is a leading resource of protein information, providing the research community with a comprehensive, high quality and freely accessible platform of protein sequences and functional information. Manual curation of a protein entry includes sequence analysis, functional information from literature, and identification of orthologs.

For human entries, we also provide information on diseases associated with genetic variations in a given protein. The information is extracted from scientific literature and diseases that are also described in the OMIM database. Understanding the association of genetic variation with its functional consequences in proteins is essential for the interpretation of genomic data and identifying causal variants in diseases. Therefore, our ongoing efforts aim to compile all available information on proteins associated with human diseases, including the annotation of protein variants, in order to help researchers to better understand the relationship between protein function and disease. We are manually annotating disease-associated variants and assess pathogenicity based on the criteria by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Information of human proteins involved in disease is supplemented with data from protein orthologs in model organisms, providing additional information based on mutagenesis assays and disruption phenotypes.

With these efforts, we hope to improve our understanding between the association of genetic variation with its functional consequences on proteins. This is essential for the interpretation of genomic data and identifying causal variants in diseases.

23. Curating Somatic Variants in Rare Head and Neck Cancers in COSMIC

CG Cole, A Holmes, A Sangrador-Vegas, M Starkey, D Carvalho-Silva, J Argasinska, R Lyne, R White, Sari A Ward
COSMIC, Wellcome Sanger Institute

Head and neck cancer is a relatively uncommon type of cancer. Around 12,400 new cases are diagnosed in the UK each year (NHS). The most common type is the squamous cell carcinoma, which develops in the squamous cells that line the mucosal surfaces of the head and neck (for example, those inside the mouth, throat, and voice box). We have focused our curation on the less common head and neck cancers, for example the ones that develop in the salivary glands, sinuses, or muscle and bone in the head and neck.

For the upcoming May release v96 of COSMIC, the Catalogue of Somatic Mutations in Cancer, we curated variant and patient related metadata from 56 head and neck cancer publications. From these publications, 129 new organ site-histology pairs were added to COSMIC. Data was added per sample and each sample was accompanied with sequencing data. In addition we have identified two new cancer genes, PRKD1 and MUC6, as defining a subset of H&N cancers. The role of these genes in cancer was evaluated and both were added to Cancer Gene Census, a manually curated resource of genes that contribute to cancer development. The literature reporting PRKD1 and MUC6 variants across all cancers was comprehensively curated and will be released in COSMIC versions 96 and 97.

In recent years there has been a rapid growth in data production and publications relating to cancer, genetics, and variants. The role of curators in distilling, organising and highlighting key data has become increasingly important. Here, we discuss COSMIC's process for selective and structured curation to extract high quality and valuable data. This includes our process to prioritise new cancer genes, rare tumours, and new tumour types not yet represented in our database, or new variants within well classified genes/tumour types. Our overarching goal is a comprehensive yet high-resolution dataset that can be meaningfully utilised from discovery at the laboratory bench to decision-making at the patient bedside.

24. Rfam, curating families with 3D structures

Nancy Ontiveros-Palacios, E Cooke, A Bateman, B A. Sweeney and A I. Petrov
EMBL-EBI

Rfam (https://rfam.org) is a comprehensive database of RNA families, each represented by a manually curated alignment, a consensus secondary structure and a covariance model. Rfam is nearly 20 years old and since its inception, it has become the largest and most widely used database of non-coding RNA (ncRNA) families, currently including over 4,000 families. Historically, the RNA secondary structure for each family was obtained through bioinformatic predictions before a 3D structure was solved. Now, the Rfam team is using available 3D information from Protein Data Bank (PDB) to review ncRNA families and provide more accurate secondary structures. As of March 2022 the 3D structures of 124 Rfam families have been reported and we are currently updating the families by adding missing base pairs, as well as other missing structural elements, including pseudoknot elements and long distance base pairs. Additionally, we are including annotations, such as RNA 3D motifs (for example k-turn), RNA structural elements (stems, hairpin loops, junctions), and ligands (particularly in riboswitches). In the 30 first reviewed families, the secondary structure of 27 families was improved by updating missing pseudoknots in 19 of the families and adding or correcting base pairs in 26 of the families. For example, the central part of the flavin mononucleotide (FMN) riboswitch is now organized by several additional base pairs and two pseudoknots. The first 30 updated families have been released and include riboswitches, coronavirus RNAs, spliceosomal RNAs, ribozymes, microRNAs, and other RNAs. Rfam is continuously improving the quality of RNA families, and this targeted review of families using 3D information fills the gap between RNA-predicted structures and the experimentally determined RNA 3D structures.

25. EMPIAR – for all your raw data archiving needs

Simone Weyand, O Salih, C Catavitello, A Iudin, P Korir, S Somasundharam, G Kleywegt, A Patwardhan
EMBL-EBI

The EMPIAR database was established in 2013 and covers several modalities including cryo-EM (with related EMDB entry), cryo-ET, SBF-SEM, FIB-SEM, CLEM/CLXM (LM data archived in BioImage Archive), ATUM-SEM, ssET, SXT, Hard X-ray/X-ray MicroCT, MicroED. It is publicly available and free of charge in order to make raw data available to the worldwide community.

List of attendees

First Name	Last Name	Email Address	Company/University Name	Talk (T)/Poster(P) Number
Yasmin	Alam-Faruque	yasmin.alam-faruque@healx.io	Healx	
Joanna	Argasinka	ja30@sanger.ac.uk	Wellcome Sanger Institute	
Helen	Attrill	hla28@cam.ac.uk	FlyBase/Cambridge University	T17, T22
Kenneth	Baillie	j.k.baillie@ed.ac.uk	University of Edinburgh	
John	Berrisford	john.berrisford@astrazeneca.com	AstraZeneca	
Kamel Eddine Adel	Bouhraoua	adelbouhraoua23@gmail.com	ELIXIR ITALY	
Emily	Bowler-Barnett	ebowler@ebi.ac.uk	EMBL-EBI	P17
Bryony	Braschi	bbraschi@ebi.ac.uk	EMBL - EBI	P2
Natassja	Bush	nbush@ebi.ac.uk	EMBL-EBI	
Nancy	Campbell	nancy.campbell@healx.io	Healx	
Elena	Cibrian-Uhalte	elena.cibrian@healx.io	Healx	
Charlotte	Cole	cgc@sanger.ac.uk	Sanger Institute	P11, P15, P23
Chuck	Cook	ccook@globalbiodata.org	Global Biodata Coalition	T7, T23
Andrea	Cosolo	andrea_cosolo@hms.harvard.edu	Harvard Medical School	
Melanie	Courtot	mcourtot@oicr.on.ca	Ontario Institute for Cancer Research	
Alice	Crowley	acrowley@ebi.ac.uk	EMBL/ EBI	
Denise	de Carvalho Silva	dd11@sanger.ac.uk	Wellcome Sanger Institute	P23
Paul	Denny	pdenny@ebi.ac.uk	EMBL-EBI	T18, P16
Stavros	Diamantakis	skd@ebi.ac.uk	EMBL/ EBI	P7
Rachel	Drysdale	rdrysdale@globalbiodata.org	Global Biodata Coalition	T23
Ibrahim	Emam	iemam@ic.ac.uk	Imperial College London	T3
Gavin	Farrell	gavin.farrell@elixir-europe.org	ELIXIR EUROPE (EBI)	
Rebecca	Foulger	rebecca@scibite.com	SciBite	

First Name	Last Name	Email Address	Company/University Name	Talk (T)/Poster(P) Number
Anja	Fullgrabe	anjaf@ebi.ac.uk	EMBL/ EBI	
Penelope	Garmiri	penelope.garmiri@healx.io	Healx	
Nancy	George	ngeorge@ebi.ac.uk	EMBL/ EBI	T21
George	Georghiou	geo.georghiou@gmail.com	Novartis	
Sucheta	Ghosh	sucheta.ghosh@h-its.org	Heidelberg Institute for Theoretical Studies	T9
Damien	Goutte-Gattat	dpg44@cam.ac.uk	University of Cambridge	P1
Pratibha	Gour	pratibha.gour@gmail.com	University of Delhi	T10
Melissa	Harrison	mharrison@ebi.ac.uk	EMBL/ EBI	
Jennifer	Harrow	jen.harrow@elixir-europe.org	ELIXIR	
Henning	Hermjakob	hhe@ebi.ac.uk	EMBL-EBI	T20, P21
Alexandra	Holinski	aholinski@ebi.ac.uk	EMBL-EBI	
Alex	Holmes	ah30@sanger.ac.uk	Sanger Institute	P11, P23
Mohammad	Hosseini	mohammad.hosseini@northwestern.edu	Northwestern University	
Chris	Hunter	chris@gigasciencejournal.com	GigaScience Press	T1, P8
Rachael	Huntley	r.huntley@elsevier.com	SciBite	
Bijay	Jassa;	bijay.jassal@astrazeneca.com	AstraZeneca	
Anneli	Karlsson	anneli@scibite.com	SciBite	
Carlo	Kroll	bt211093@qmul.ac.uk	Queen Mary University of London	T6, P10
Deepti Jaiswal	Kundu	jaiswal@ebi.ac.uk	EMBL/ EBI	T16
Manuel	Lera Ramirez	manulera14@gmail.com	UCL	P3
Jake	Lever	jake.lever@glasgow.ac.uk	University of Glasgow	T11
Sonia	Liggi	sonia.liggi@healx.io	Healx	
Antonia	Lock	alock@ebi.ac.uk	UniProt / Embl-EBI	P18
Jane	Lomax	jane@scibite.com	SciBite	
Ruth	Lovering	r.lovering@ucl.ac.uk	University College London	T13, T22, P14

First Name	Last Name	Email Address	Company/University Name	Talk (T)/Poster(P) Number
Yvonne	Lussi	ylussi@ebi.ac.uk	EMBL/ EBI	T18, P22
Rachel	Lyne	rl24@sanger.ac.uk	Wellcome Sanger Institute	P23
Michele	Magrane	magrane@ebi.ac.uk	EMBL - EBI	T18, P16, P20, P22
Paloma	Marín-Arraiza	p.arraiza@orcid.org	ORCID	
Maria	Martin	martin@ebi.ac.uk	EMBL/ EBI	T8
Steven	Marygold	sjm41@cam.ac.uk	FlyBase, University of Cambridge	
Stacy	Mather	stacy.mather@astrazeneca.com	AstraZeneca	
Alex	McLachlan	adm71@cam.ac.uk	University of Cambridge	P1
Peter	McQuilton	peter.x.mcquilton@gsk.com	GSK	
Birgit	Meldal	birgit.meldal@eaglegenomics.com	Eagle Genomics Ltd	
Sarah	Morgan	sarahm@ebi.ac.uk	EMBL-EBI	
Nancy	Ontiveros	nancyontiveros@ebi.ac.uk	EMBL-EBI	T19, P24
Sandra	Orchard	orchard@ebi.ac.uk	EMBL-EBI	T18, T20, P16, P17, P18, P20, P21, P22
Arzu	Ozturk Colak	ao493@cam.ac.uk	FlyBase, University of Cambridge	
Kalpana	Panneerselvam	kalpanap@ebi.ac.uk	EMBL-EBI	T20, P21
Joana	Pauperio	Joanap@ebi.ac.uk	EMBL/ EBI	T14
Diego	Pava	bt211047@qmul.ac.uk	Queen Mary University of London	T6, P10
Thomas	Payne	payne@ebi.ac.uk	EMBL-EBI	
Livia	Perfetto	livia.perfetto@fht.org	Fondazione Human Technopole	T20, P21
Clare	Pilgrim	cp390@cam.ac.uk	University of Cambridge	P1
Damiano	Piovesan	damiano.piovesan@unipd.it	University of Padova	T5
Pablo	Porras Millan	pablo.porrasmillan@astrazeneca.com	AstraZeneca	T20, P21
Arina	Puzriakova	arina.puzriakova@genomicsengland.co.uk	Genomics England	
Federica	Quaglia	federica.quaglia8@gmail.com	National Research Council (Italy)	T13

First Name	Last Name	Email Address	Company/University Name	Talk (T)/Poster(P) Number
Pedro	Raposorap	raposo@ebi.ac.uk	EMBL-EBI	P9
Colbie	Reed	creed212@ufl.edu	University of Florida	P12
Vasileios	Sagris	sagris.vasileios@ucy.ac.cy	University of Cyprus	
Amaia	Sangrador Vegas	as52@sanger.ac.uk	Sanger Institute	P11, P23
Enrique	Sapena Ventura	enrique@ebi.ac.uk	EMBL/ EBI	T2, P5
Shirin	Saverimuttu	shirin@scibite.com	SciBite	
Catherine	Snow	catherine.snow@genomicsengland.co.uk	Genomics England	
Elena	Speretta	esperett@ebi.ac.uk	EMBL/ EBI	P16, P22
Michaela	Spitzer	mspitzer@exscientia.co.uk	Exscientia	
Michael	Starkey	ms66@sanger.ac.uk	Sanger Institute	P15, P23
Elisa	Stefaniak	elisa@foxeswithdata.com	Foxes With Data	
James	Stephenson	jstephenson@ebi.ac.uk	EMBL-EBI	T8
Anna	Swan	annas@ebi.ac.uk	EMBL-EBI	
Blake	Sweeney	bsweeney@ebi.ac.uk	EMBL-EBI	T19, P24
Wei Kheng	Teh	Wteh@ebi.ac.uk	EMBL-EBI	
Mary-Ann	Tuli	maryann@gigasciencejournal.com	GigaScience Press	T1, T13, T24
Nidhi	Tyagi	tyagi@ebi.ac.uk	EMBL-EBI	
Sameer	Velankar	sameer@ebi.ac.uk	EMBL/ EBI	
Aravind	Venkatesan	avenkat@ebi.ac.uk	EMBL-EBI	
Liora	Vilmovsky	liora@ebi.ac.uk	EMBL-EBI	P2
Sari	Ward	sw13@sanger.ac.uk	Sanger Institute	P11, P15, P23
Simone	Weyand	simonew@ebi.ac.uk	EMBL-EBI	P25
Rebecca	White	rw19@sanger.ac.uk	Sanger Institute	P23
Eleanor	Williams	eleanor.williams@genomicsengland.co.uk	Genomics England	
Ulrike	Wittig	ulrike.wittig@h-its.org	Heidelberg Institute for Theoretical Studies	T9

First Name	Last Name	Email Address	Company/University Name	Talk (T)/Poster(P) Number
Valerie	Wood	vw253@cam.ac.uk	University of Cambridge	T15
Yannick	Wurm	y.wurm@qmul.ac.uk	Queen Mary University of London	T6, P10
Magdalena	Zarowiecki	mz3@ebi.ac.uk	EMBL/ EBI	P7
Rossana	Zaru	rzaru@ebi.ac.uk	EMBL/ EBI	P16, P20